Ciberseguridad y Seguridad del Paciente: Una Revisión Sistemática sobre las Vulnerabilidades de la Inteligencia Artificial en el Diagnóstico por Imágenes Médicas.

María Valentina Carvajal Zuluaga.

Asesor

Edna Rocío Jamaica Guio

Universidad Nacional Abierta y a Distancia - UNAD

Escuela de Ciencias de la Salud - ECISA

Tecnología en Radiología e Imágenes Diagnosticas

2026

Agradecimientos.

La realización de este trabajo representa un paso importante dentro de mi proceso de formación académica y personal y por ello, en primer lugar, agradezco a Dios por brindarme la fortaleza, la sabiduría y la oportunidad de avanzar en este camino de aprendizaje, permitiéndome superar cada reto presentado durante el desarrollo de este proyecto.

Expreso también mi profundo agradecimiento a mi familia, quienes han sido un apoyo fundamental a lo largo de este proceso, su amor, comprensión, confianza y motivación constante han sido un impulso para continuar adelante y no rendirme ante las dificultades.

De manera especial, agradezco a mi directora y asesora Edna Rocío Jamaica Guio por su dedicación, compromiso y orientación durante el desarrollo de este trabajo, ya que sus conocimientos, acompañamiento y disposición para guiar cada etapa del proceso académico fueron esenciales para fortalecer mi aprendizaje y alcanzar los objetivos propuestos.

Asimismo, agradezco a mis compañeros de carrera, con quienes compartí experiencias, conocimientos y aprendizajes que contribuyeron significativamente a mi formación académica y personal.

Finalmente, expreso mi sincero agradecimiento a la Universidad Nacional Abierta y a Distancia, institución que me ha brindado la oportunidad de formarme profesionalmente, permitiéndome desarrollar nuevas habilidades y continuar creciendo en mi proyecto de vida

Resumen.

La incorporación de la inteligencia artificial en el diagnóstico por imágenes médicas ha generado un cambio significativo en la práctica radiológica, permitiendo mejorar la detección precoz de enfermedades y aumentar la precisión en la interpretación de estudios, y, sin embargo, este avance también plantea retos importantes relacionados con la seguridad, la confiabilidad y el uso adecuado de estas tecnologías en contextos clínicos. El presente trabajo se fundamenta en una revisión sistemática de la literatura, con el propósito de identificar y analizar las principales vulnerabilidades de los sistemas de inteligencia artificial aplicados a la radiología, así como su impacto en la seguridad del paciente.

Mediante un enfoque cualitativo y un diseño de tipo documental, se revisó evidencia científica publicada entre 2019 y 2025, junto con normativas y lineamientos internacionales donde los resultados evidencian la presencia de riesgos emergentes, entre ellos los ataques adversariales y la alteración de imágenes mediante técnicas avanzadas, que pueden generar errores diagnósticos difíciles de detectar, incluyendo falsos positivos y negativos. De igual forma, se identificó que la limitada transparencia de algunos algoritmos, conocida como efecto de “caja negra”, dificulta la comprensión de los procesos de decisión, lo que puede afectar la confianza del personal de salud.

Dando finalidad, se resalta la importancia de fortalecer la ciberseguridad y promover sistemas más transparentes y explicables. Asimismo, se hace énfasis en la necesidad de establecer regulaciones sólidas y mantener una supervisión humana constante, con el fin de garantizar un uso seguro, ético y responsable de la inteligencia artificial en la atención en salud.

Palabras Clave: Inteligencia artificial, ciberseguridad en salud, diagnóstico por imágenes, ataques adversariales, seguridad del paciente, radiología digital

Abstract

The incorporation of artificial intelligence into medical imaging diagnosis has generated a significant change in radiological practice, allowing for improved early detection of diseases and increased accuracy in the interpretation of studies; however, this advancement also poses important challenges related to security, reliability, and the proper use of these technologies in clinical contexts. This work is based on a systematic review of the literature, with the purpose of identifying and analyzing the main vulnerabilities of artificial intelligence systems applied to radiology, as well as their impact on patient safety.

Through a qualitative approach and a documentary research design, scientific evidence published between 2019 and 2025 was reviewed, along with international regulations and guidelines. The results reveal the presence of emerging risks, including adversarial attacks and image manipulation through advanced techniques, which can generate diagnostic errors that are difficult to detect, including false positives and false negatives. Likewise, it was identified that the limited transparency of some algorithms, known as the "black box" effect, hinders the understanding of decision-making processes, which may affect the confidence of healthcare personnel.

In conclusion, the importance of strengthening cybersecurity and promoting more transparent and explainable systems is highlighted. Likewise, emphasis is placed on the need to establish robust regulations and maintain constant human oversight in order to ensure the safe, ethical, and responsible use of artificial intelligence in healthcare.

Keywords: Artificial intelligence, cybersecurity in healthcare, medical imaging diagnosis, adversarial attacks, patient safety, digital radiology.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

En la actualidad, los avances tecnológicos han impulsado una transformación significativa en el campo de la salud, destacándose especialmente la incorporación de la inteligencia artificial (IA) en el análisis y procesamiento de imágenes médicas. Estas tecnologías, basadas principalmente en algoritmos de aprendizaje profundo, han demostrado una alta capacidad para identificar patrones complejos, lo que ha contribuido al fortalecimiento de los procesos diagnósticos, facilitando la detección temprana de diversas enfermedades y optimizando la toma de decisiones clínicas por parte de los profesionales de la salud. En este sentido, a partir de una revisión documental y análisis de la evidencia científica reciente, diversos estudios han demostrado que la IA puede alcanzar niveles de precisión comparables, e incluso superiores, a los de expertos humanos en áreas como la radiología, la dermatología y la oftalmología, lo que resalta su potencial como herramienta de apoyo en la práctica clínica (Topol, 2019; Esteva et al., 2021).

No obstante, a pesar de los beneficios asociados a la implementación de estos sistemas, también emergen nuevos desafíos que impactan directamente la seguridad del paciente y la calidad de la atención. Desde un análisis basado en la evidencia, la confiabilidad de los algoritmos, la presencia de sesgos en los modelos de aprendizaje automático y la falta de transparencia en su funcionamiento representan preocupaciones relevantes en la actualidad. Investigaciones recientes han señalado que los sistemas de IA pueden reproducir e incluso amplificar desigualdades existentes en los datos de entrenamiento, lo que puede traducirse en errores diagnósticos o decisiones clínicas inadecuadas (Rajkomar et al., 2019; Kelly et al., 2021). A su vez, estos riesgos se ven potenciados por la complejidad inherente al razonamiento clínico, el cual, como se ha descrito previamente, no es completamente lineal y puede estar influenciado

por sesgos cognitivos que afectan la interpretación de los resultados proporcionados por estas tecnologías.

Adicionalmente, el uso creciente de sistemas digitales en salud ha incrementado la exposición a amenazas relacionadas con la ciberseguridad. A partir de la revisión de literatura científica especializada, se evidencia que la manipulación de imágenes médicas mediante ataques adversariales, así como el acceso no autorizado a datos sensibles, constituyen riesgos emergentes que pueden comprometer la integridad de la información clínica y afectar la precisión diagnóstica. Estudios recientes han demostrado que pequeñas modificaciones imperceptibles en imágenes médicas pueden alterar significativamente las predicciones de los modelos de IA, generando falsos positivos o falsos negativos con potencial impacto en la seguridad del paciente (Finlayson et al., 2019; Paschali et al., 2021). Estas vulnerabilidades ponen de manifiesto la necesidad de fortalecer los sistemas de protección y desarrollar mecanismos robustos que garanticen la confiabilidad de estas herramientas tecnológicas.

En este contexto, el presente trabajo se desarrolla bajo un enfoque de revisión documental y análisis crítico de la evidencia, con el propósito de abordar los principales conceptos y fundamentos relacionados con la inteligencia artificial aplicada al diagnóstico por imágenes, la ciberseguridad en los sistemas de información en salud y las vulnerabilidades que pueden afectar el funcionamiento de estas tecnologías. Asimismo, se pretende analizar la importancia de garantizar la protección de los datos médicos, la transparencia algorítmica y la integridad de los sistemas tecnológicos utilizados en el ámbito sanitario, considerando tanto los factores tecnológicos como los humanos implicados en el proceso diagnóstico.

De esta manera, el desarrollo del marco teórico permitirá comprender de manera integral la relación existente entre la innovación tecnológica, el razonamiento clínico, la seguridad de la

información y la protección del paciente. Todo ello resalta la necesidad de implementar estrategias multidisciplinarias, sustentadas en la evidencia científica, que promuevan un uso responsable, ético y seguro de la inteligencia artificial en el sector salud, orientado a minimizar riesgos, reducir errores diagnósticos y mejorar la calidad de la atención sanitaria en contextos cada vez más digitalizados.

Planteamiento del Problema

El avance de las tecnologías digitales ha impulsado el desarrollo y la implementación de sistemas de inteligencia artificial en diferentes áreas del sector salud, especialmente en el diagnóstico por imágenes médicas. Estas herramientas han permitido optimizar el análisis de estudios radiológicos, mejorar la detección temprana de enfermedades y apoyar la toma de decisiones clínicas. No obstante, a pesar de sus múltiples beneficios, el uso de estos sistemas también ha generado preocupaciones relacionadas con la seguridad, confiabilidad y protección de la información médica.

Uno de los principales problemas asociados con la implementación de la inteligencia artificial en el diagnóstico por imágenes se relaciona con la existencia de vulnerabilidades en los sistemas informáticos que pueden ser explotadas mediante ataques cibernéticos o manipulaciones digitales. Entre estos riesgos se encuentran los ataques adversariales, que consisten en la alteración intencional de imágenes médicas para modificar los resultados generados por los algoritmos de inteligencia artificial, lo que puede derivar en interpretaciones erróneas y afectar la precisión diagnóstica.

Asimismo, los modelos de inteligencia artificial pueden presentar limitaciones relacionadas con la presencia de sesgos en los datos utilizados para su entrenamiento, lo que puede provocar errores en la identificación de patologías, generando falsos positivos o falsos negativos. Estas situaciones pueden influir directamente en la toma de decisiones médicas, comprometiendo la seguridad del paciente y la confiabilidad de los sistemas tecnológicos aplicados en el ámbito de la salud.

En este contexto, surge la necesidad de analizar las vulnerabilidades existentes en los sistemas de inteligencia artificial aplicados al diagnóstico por imágenes médicas, así como su

impacto en la seguridad del paciente y en la protección de los datos clínicos. Comprender estas problemáticas permitirá identificar los desafíos actuales relacionados con la ciberseguridad en salud y promover el desarrollo de estrategias que garanticen un uso seguro, ético y confiable de estas tecnologías en los procesos de atención médica.

Justificación.

En los últimos años, el desarrollo tecnológico ha permitido la incorporación de herramientas innovadoras en el campo de la salud, entre las cuales destaca la inteligencia artificial aplicada al análisis de imágenes médicas, y estas tecnologías han contribuido significativamente a mejorar los procesos de diagnóstico, permitiendo una mayor rapidez en la detección de diversas patologías, así como una mayor precisión en la interpretación de estudios radiológicos, donde, gracias a estos avances, los profesionales de la salud cuentan con sistemas que apoyan la toma de decisiones clínicas y optimizan la calidad de la atención brindada a los pacientes (Topol, 2019; Hosny et al., 2018).

No obstante, a pesar de los beneficios que ofrece la implementación de la inteligencia artificial en el ámbito médico, su uso también ha puesto en evidencia ciertos desafíos relacionados con la seguridad y confiabilidad de los sistemas tecnológicos y el creciente uso de plataformas digitales para el almacenamiento, procesamiento y análisis de datos médicos ha incrementado el riesgo de vulnerabilidades informáticas que podrían comprometer tanto la integridad de la información clínica como la precisión de los resultados diagnósticos (Kelly et al., 2019; WHO, 2021).

Entre las principales preocupaciones se encuentran los ataques adversariales, la manipulación de imágenes médicas y las posibles fallas en los algoritmos de aprendizaje automático. Estas situaciones pueden alterar el funcionamiento de los sistemas de inteligencia artificial, generando interpretaciones incorrectas como falsos positivos o falsos negativos en los diagnósticos, y como consecuencia, estas fallas podrían influir en la toma de decisiones médicas, afectar el tratamiento de los pacientes e incluso representar un riesgo para su seguridad (Finlayson et al., 2019; Paschali et al., 2021).

En este contexto, resulta fundamental analizar las vulnerabilidades presentes en los sistemas de inteligencia artificial aplicados al diagnóstico por imágenes, así como comprender la importancia de implementar medidas de ciberseguridad que permitan proteger los datos médicos y garantizar la confiabilidad de los sistemas tecnológicos utilizados en el sector salud, además del fortalecimiento de la seguridad informática en estos sistemas no solo contribuye a preservar la integridad de la información clínica, sino que también favorece la confianza de los profesionales de la salud y de los pacientes en el uso de estas tecnologías (European Commission, 2021; FDA, 2019).

Por lo tanto, este trabajo se justifica en la necesidad de reflexionar sobre la relación existente entre inteligencia artificial, diagnóstico por imágenes médicas y ciberseguridad, con el propósito de identificar los principales riesgos y desafíos asociados a su implementación, asimismo, busca generar mayor conciencia sobre la importancia de desarrollar estrategias y mecanismos de protección que permitan garantizar un uso responsable, seguro y confiable de la inteligencia artificial en el ámbito de la radiología y en la atención médica en general (WHO, 2021; Topol, 2019).

Objetivos

Objetivo General

Analizar, a partir de la revisión documental y la evidencia científica disponible, las vulnerabilidades de los sistemas de inteligencia artificial aplicados al diagnóstico por imágenes médicas y su impacto en la seguridad del paciente, destacando el papel fundamental de la ciberseguridad en el ámbito de la salud.

Objetivos Específicos

Identificar, a partir de la revisión documental, las principales aplicaciones de la inteligencia artificial en el diagnóstico por imágenes médicas, analizando su evolución y uso en la práctica clínica.

Analizar, con base en la literatura científica, los riesgos y vulnerabilidades asociados al uso de sistemas de inteligencia artificial en radiología, incluyendo los ataques adversariales y la manipulación de imágenes médicas.

Explicar, a partir de la evidencia revisada, el impacto de los errores diagnósticos, especialmente los falsos positivos y falsos negativos, en los sistemas de diagnóstico asistido por inteligencia artificial.

Describir, analizando la literatura especializada, la importancia de la ciberseguridad en la protección de los datos médicos y en la confiabilidad de los sistemas de inteligencia artificial en salud.

Reflexionar, a partir de la evidencia científica y los marcos regulatorios revisados, acerca de la necesidad de implementar estrategias de seguridad y regulación tecnológica que garanticen un uso ético, seguro y responsable de la inteligencia artificial en el ámbito sanitario

Marco Teórico.

Conceptualización de la Inteligencia Artificial en el Acto Médico

La incorporación de la inteligencia artificial (IA) en el ámbito sanitario representa una transformación estructural del acto médico, particularmente en el proceso diagnóstico, en términos generales, la IA aplicada a la medicina puede entenderse como el conjunto de sistemas computacionales capaces de analizar grandes volúmenes de datos clínicos, identificar patrones complejos y generar inferencias que contribuyen a la toma de decisiones médicas, sin embargo, su relevancia no radica únicamente en la capacidad técnica de procesamiento de datos, sino en el grado de autonomía decisional que puede asumir dentro del proceso asistencial (Topol, 2019).

En el acto médico tradicional, el diagnóstico es el resultado de un proceso cognitivo desarrollado por el profesional de la salud, quien integra información clínica, antecedentes, hallazgos físicos y pruebas complementarias para formular una hipótesis explicativa del estado del paciente, ya que la introducción de sistemas de IA altera esta dinámica al incorporar un agente tecnológico capaz de participar activamente en la formulación diagnóstica, y esta participación puede variar desde un rol meramente auxiliar hasta una función decisoria con mínima intervención humana, lo que obliga a replantear categorías clásicas como responsabilidad profesional, error clínico y estándar de diligencia (Benjamens et al., 2020).

Desde una perspectiva conceptual, la IA en medicina no debe entenderse únicamente como una herramienta informática avanzada, sino como una tecnología con capacidad de incidir directamente en la seguridad del paciente, en la estructura del razonamiento clínico y en la organización del sistema sanitario, y por ello, la doctrina científica ha comenzado a analizarla no solo desde el enfoque técnico, sino también ético y jurídico, especialmente cuando su aplicación implica decisiones que afectan de manera directa la salud o la vida del paciente (Topol, 2019).

Figura 1

La incorporación de la inteligencia artificial (IA) en el ámbito sanitario

Nota. Uno de los avances más significativos en imágenes médicas es el uso de la inteligencia artificial para el diagnóstico y la interpretación. Fuente. Revistahospitalaria.org. (2025)

Definición de Sistemas Autónomos de IA: Diferencia entre Algoritmos Pasivos y Sistemas Activos

Una distinción fundamental dentro del estudio de la IA médica radica en diferenciar los algoritmos de apoyo a la decisión clínica de los sistemas autónomos de IA, y esta diferenciación no es meramente terminológica, sino que tiene consecuencias relevantes en términos de supervisión humana, validación clínica y responsabilidad (European Commission, 2021). Los sistemas de apoyo a la decisión clínica (Clinical Decision Support Systems, CDSS) operan bajo un modelo denominado *human-in-the-loop*, en el cual el algoritmo procesa información como imágenes médicas, datos bioquímicos o registros electrónicos y genera recomendaciones o probabilidades diagnósticas que deben ser interpretadas y validadas por el profesional médico, antes de adoptar una decisión final (Jiang et al., 2017; Sutton et al., 2020). , en este modelo, la IA cumple una función instrumental que complementa el juicio clínico, pero no lo sustituye.

En contraste, los sistemas autónomos de IA pueden emitir decisiones diagnósticas o terapéuticas sin intervención directa del profesional en el momento de la ejecución, en este esquema, conocido como *human-on-the-loop*, el médico supervisa el funcionamiento general del sistema, pero no participa activamente en cada decisión individual (He et al., 2019). Este desplazamiento parcial del juicio clínico hacia el algoritmo implica una transferencia de agencia tecnológica que modifica la naturaleza del acto médico, ya que la diferencia sustancial entre ambos modelos reside en el grado de control humano sobre la decisión final, mientras los algoritmos pasivos actúan como herramientas cognitivas que amplían la capacidad de análisis del médico, los sistemas autónomos asumen funciones tradicionalmente reservadas al profesional, incluyendo la clasificación diagnóstica directa (Topol, 2019). Esta delegación funcional genera interrogantes en torno a la imputabilidad del error, especialmente cuando el desempeño algorítmico depende de variables como la calidad del entrenamiento, la representatividad de los datos o la capacidad de generalización del modelo (Benjamens et al., 2020).

El Proceso de Diagnóstico Médico y la Inserción de la IA

El diagnóstico médico es un proceso complejo que involucra múltiples niveles de razonamiento clínico, en los cuales el profesional de la salud combina tanto respuestas rápidas basadas en la experiencia previa como análisis más estructurados y reflexivos. A partir de la evidencia reciente, se reconoce que esta interacción entre pensamiento intuitivo y analítico permite una toma de decisiones más eficiente, pero también expone al clínico a posibles errores derivados de sesgos cognitivos y limitaciones en el procesamiento de la información. En este sentido, estudios actuales destacan que la precisión diagnóstica depende no solo del conocimiento clínico, sino también de la capacidad de integrar datos, interpretar imágenes y manejar la incertidumbre en contextos complejos (Norman et al., 2021).

La IA se inserta principalmente en aquellas fases del proceso diagnóstico relacionadas con el reconocimiento de patrones y el análisis de grandes volúmenes de información, en especialidades como la radiología y la patología digital, los modelos de aprendizaje profundo han demostrado alta capacidad para identificar anomalías en imágenes médicas, tales como lesiones tumorales o alteraciones histológicas (Hosny et al., 2018), en estos contextos se evidencia que la IA puede actuar como herramienta de apoyo o como sistema autónomo capaz de emitir una clasificación diagnóstica y en el ámbito de la patología digital, los algoritmos basados en *deep learning* pueden analizar láminas histológicas digitalizadas y detectar patrones microscópicos con niveles de precisión comparables a los de expertos humanos (Campanella et al., 2019). No obstante, cuando estos sistemas operan sin revisión inmediata por parte del especialista, el estándar de validación y supervisión debe ser más estricto, dado que el margen de error puede afectar directamente la seguridad del paciente. Así, la inserción de la IA en el proceso diagnóstico no sustituye necesariamente el razonamiento clínico humano, pero sí reconfigura su estructura, generando un modelo híbrido en el cual la interacción entre profesional y algoritmo determina el resultado final.

La interpretación de imágenes diagnósticas constituye una actividad fundamental dentro del proceso clínico, ya que permite identificar patologías y orientar decisiones terapéuticas, donde tradicionalmente, esta labor ha sido realizada por radiólogos especializados mediante la evaluación directa de estudios como radiografías, tomografías computarizadas y resonancias magnéticas, sin embargo, el avance de la inteligencia artificial (IA), particularmente de los modelos de aprendizaje profundo aplicados a la visión computacional, ha permitido el desarrollo de herramientas capaces de analizar grandes volúmenes de imágenes médicas con altos niveles de precisión (Topol, 2019).

En este contexto, el surgimiento de la teleradiología ha facilitado la transmisión digital de imágenes médicas para su interpretación remota, lo cual ha incrementado la eficiencia en la prestación de servicios diagnósticos. Paralelamente, los sistemas de IA han comenzado a integrarse como herramientas de apoyo en este proceso, generando un nuevo modelo de análisis que combina la experiencia humana con la capacidad computacional de los algoritmos (Hosny, Parmar, Quackenbush, Schwartz & Aerts, 2018). No obstante, esta transformación tecnológica también introduce nuevos retos relacionados con la seguridad del paciente, la confiabilidad de los algoritmos, la ciberseguridad de los sistemas y la regulación del error diagnóstico, aspectos que actualmente son objeto de análisis por parte de organismos regulatorios como la Food and Drug Administration y la Unión Europea.

Figura 2

Flujo de la teleradiologia

Nota. Los flujos de trabajo en teleradiología apoyados con IA permiten agilizar el tiempo de los reportes y son fundamentales para disminuir el cansancio y la fatiga en el profesional. Fuente. Aporte realizado por la docente Edna Rocío Jamaica Guio

Comparación Entre el Proceso Tradicional y el Proceso Asistido por Inteligencia Artificial

Proceso Tradicional de Interpretación Radiológica

En el modelo tradicional, el diagnóstico por imágenes depende principalmente de la experiencia clínica del radiólogo donde, el especialista revisa manualmente las imágenes médicas generadas por diferentes equipos de diagnóstico y, mediante su conocimiento anatómico y patológico, identifica anomalías o patrones que puedan indicar la presencia de enfermedad (Brady, 2017). Este proceso incluye varias etapas: adquisición de la imagen, procesamiento básico, revisión por el especialista y elaboración del informe diagnóstico, en los cuales, en muchos casos, especialmente en instituciones con limitaciones de personal médico, la interpretación puede retrasarse debido al alto volumen de estudios radiológicos.

En este contexto surge la teleradiología, que permite enviar imágenes médicas a especialistas ubicados en diferentes lugares para su análisis remoto. Esta modalidad ha mejorado el acceso a servicios diagnósticos, especialmente en regiones con escasez de radiólogos, aunque sigue dependiendo completamente de la interpretación humana (Silva, Breslau & Barr, 2018).

Proceso Asistido por Inteligencia Artificial

El modelo asistido por inteligencia artificial introduce sistemas algorítmicos capaces de analizar imágenes médicas utilizando técnicas de aprendizaje automático y redes neuronales profundas, y estos sistemas pueden identificar patrones complejos dentro de grandes conjuntos de datos, lo que les permite detectar signos tempranos de enfermedades con un alto grado de precisión (Hosny et al., 2018). En este modelo, el algoritmo procesa automáticamente las imágenes médicas, identifica posibles anomalías y genera alertas o recomendaciones diagnósticas que posteriormente son revisadas por el radiólogo. En lugar de reemplazar al

profesional, la IA funciona principalmente como un sistema de apoyo que ayuda a priorizar casos críticos y a mejorar la eficiencia en la interpretación (Topol, 2019).

Sin embargo, la implementación de IA en radiología también ha evidenciado ciertos riesgos. Los algoritmos pueden cometer errores si los datos utilizados para su entrenamiento presentan sesgos o si las imágenes son manipuladas digitalmente mediante ataques adversariales. Estas vulnerabilidades pueden provocar interpretaciones incorrectas, lo que plantea importantes preocupaciones sobre la seguridad del paciente y la integridad de los datos clínicos (Finlayson et al., 2019).

Tabla 1

Flujo de la teleradiologia

Aspecto	Proceso tradicional (interpretación radiológica convencional)	Proceso asistido por Inteligencia Artificial (IA)	Fuente (Entidad/Publicación)
Responsable principal del diagnóstico	El diagnóstico es realizado exclusivamente por el radiólogo, quien analiza las imágenes médicas basándose en su experiencia clínica y conocimientos especializados.	El sistema de IA analiza inicialmente las imágenes médicas mediante algoritmos de aprendizaje automático y posteriormente el radiólogo revisa y valida los resultados generados por el sistema.	Sociedad Española de Radiología Médica (SERAM): Informe sobre la integración de la IA en el flujo de trabajo radiológico.
Método de análisis de imágenes	Interpretación visual directa realizada por el especialista mediante observación detallada de las imágenes diagnósticas.	Análisis automatizado mediante redes neuronales profundas capaces de identificar patrones complejos en grandes volúmenes de datos médicos.	The Lancet Digital Health: "Clinical validation of AI for diagnostic imaging".
Velocidad de procesamiento	El análisis depende del tiempo disponible del	Los algoritmos pueden analizar grandes	American College of Radiology (ACR): Data

Aspecto	Proceso tradicional (interpretación radiológica convencional)	Proceso asistido por Inteligencia Artificial (IA)	Fuente (Entidad/Publicación)
	radiólogo y del volumen de estudios que deba revisar y en instituciones con alta demanda puede haber retrasos diagnósticos.	cantidades de imágenes en pocos segundos, permitiendo priorizar casos urgentes y reducir tiempos de respuesta diagnóstica.	Science Institute - Workflow Efficiency Standards.
Precisión diagnóstica	Depende principalmente de la experiencia, formación y condiciones laborales del radiólogo, donde, los factores como la fatiga o la sobrecarga de trabajo pueden afectar la precisión diagnóstica.	Puede mejorar la detección de patrones complejos y anomalías sutiles; sin embargo, la precisión depende de la calidad de los datos de entrenamiento y del diseño del algoritmo.	Journal of the American Medical Association (JAMA): Comparative studies on AI vs. Human Performance.
Riesgo de error diagnóstico	Puede originarse por factores humanos como cansancio, presión laboral o interpretación subjetiva de la imagen.	Puede derivarse de sesgos algorítmicos, errores en los datos de entrenamiento, falta de actualización del modelo o manipulación adversarial de las imágenes.	Organización Mundial de la Salud (OMS): "Ethics and governance of artificial intelligence for health".
Teleradiología	Permite la transmisión digital de imágenes médicas para que radiólogos ubicados en diferentes lugares puedan interpretarlas remotamente.	La IA puede integrarse en plataformas de teleradiología para realizar un preanálisis automático antes de que el especialista revise la imagen.	Journal of Digital Imaging (SIIM): "The Role of AI in Transforming Teleradiology Workflows".
Ciberseguridad	Los riesgos se centran en la protección de los datos clínicos y en el acceso no autorizado a los sistemas hospitalarios.	Existen riesgos adicionales como ataques adversariales a los algoritmos, manipulación de	NIST (National Institute of Standards and Technology): "Adversarial Machine Learning in Healthcare".

Aspecto	Proceso tradicional (interpretación radiológica convencional)	Proceso asistido por Inteligencia Artificial (IA)	Fuente (Entidad/Publicación)
		imágenes diagnósticas o alteración de resultados generados por IA.	
Regulación	Se regula bajo las normativas tradicionales de práctica médica y dispositivos de diagnóstico por imagen.	Se regula como software médico basado en IA y es supervisado por organismos regulatorios internacionales que establecen requisitos de seguridad, transparencia y evaluación continua.	FDA (U.S. Food and Drug Administration): "Regulatory Framework for AI/ML-Based Medical Device Software".
Rol del profesional de la salud	El radiólogo tiene el control completo del proceso diagnóstico y la responsabilidad directa del informe médico.	El radiólogo mantiene la responsabilidad clínica final, pero utiliza la IA como herramienta de apoyo para mejorar la eficiencia y precisión del diagnóstico.	American College of Radiology (ACR): "Ethics of Artificial Intelligence in Radiology: Summary of the Joint Statement".
Impacto en la atención médica	Puede presentar limitaciones en contextos con escasez de especialistas o alta demanda de estudios diagnósticos.	Puede optimizar la detección temprana de enfermedades, mejorar la eficiencia del sistema de salud y apoyar la toma de decisiones clínicas.	OMS (Organización Mundial de la Salud): "Generating evidence for artificial intelligence-based medical devices".

Nota. La inteligencia artificial actúa como un asistente en la interpretación de las imágenes, apoyando al radiólogo en su labor diaria. Elaboración propia

Naturaleza de la "Caja Negra": La Falta de Explicabilidad como Raíz del Error Técnico

Uno de los principales desafíos asociados a los sistemas de inteligencia artificial en el ámbito clínico es su naturaleza de "caja negra", especialmente en modelos basados en

aprendizaje profundo. Este concepto hace referencia a la dificultad para comprender cómo estos algoritmos generan sus decisiones, a pesar de alcanzar altos niveles de precisión. A partir de la

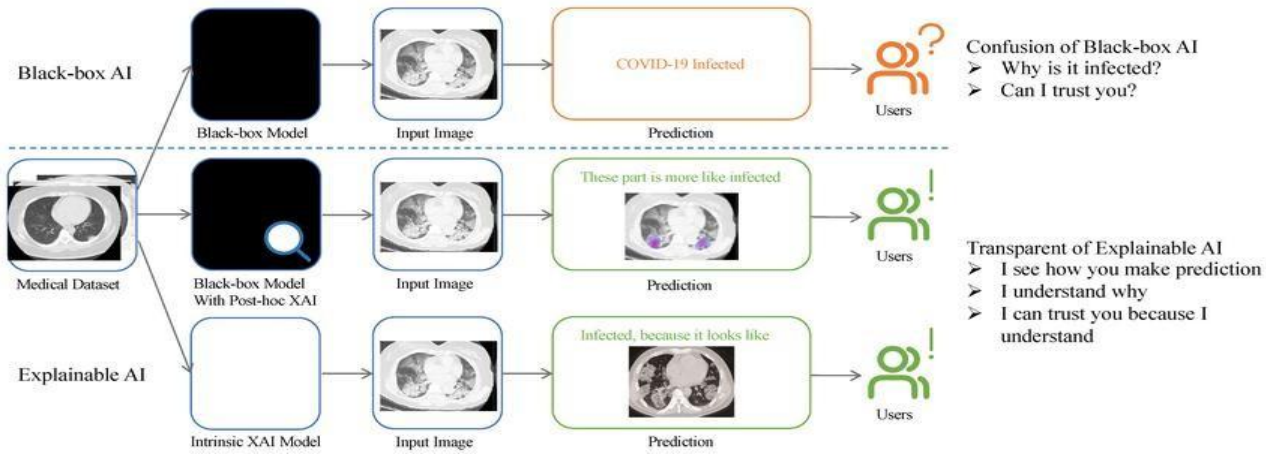
evidencia reciente, se ha señalado que la falta de interpretabilidad limita la confianza de los profesionales de la salud, dificulta la validación clínica y plantea importantes retos éticos y regulatorios. En este sentido, la literatura actual enfatiza la necesidad de desarrollar modelos explicables que permitan entender el proceso de toma de decisiones, particularmente en contextos de alto riesgo como el diagnóstico médico (Rudin, 2019; Samek et al., 2021). Aunque desde el punto de vista matemático es posible examinar los parámetros internos del modelo, la complejidad de sus múltiples capas neuronales impide traducir esa información en una justificación clínicamente inteligible. Lipton (2018) distingue entre interpretabilidad general del modelo y explicabilidad de una predicción concreta, siendo esta última especialmente relevante en el ámbito médico, donde las decisiones deben poder justificarse ante el paciente y ante instancias regulatorias.

La falta de explicabilidad no constituye únicamente una limitación epistemológica, sino una fuente potencial de error técnico, y los modelos pueden aprender correlaciones espurias presentes en los datos de entrenamiento, basando sus predicciones en patrones no causales (Zech et al., 2018). Asimismo, pueden ser vulnerables a alteraciones mínimas en los datos de entrada que modifiquen el resultado diagnóstico sin que el cambio sea perceptible para el ojo humano (Finlayson et al., 2019). Esta opacidad dificulta la auditoría del sistema y la atribución de responsabilidades en caso de daño clínico (Rudin, 2019), y desde una perspectiva ética y jurídica, la explicabilidad se vincula con principios fundamentales como la autonomía del paciente y la rendición de cuentas profesional, y por ello, el desarrollo de marcos regulatorios para la IA médica ha enfatizado la necesidad de transparencia, trazabilidad y supervisión humana significativa, especialmente cuando se trata de sistemas que influyen directamente en decisiones clínicas (European Commission, 2021).

La “caja negra” se utiliza para describir aquellos modelos algorítmicos cuyo proceso interno de toma de decisiones no es fácilmente comprensible para los seres humanos, y aunque el sistema puede generar resultados precisos, el mecanismo exacto mediante el cual llega a una conclusión diagnóstica permanece opaco o difícil de interpretar (Castelvecchi, 2016). En aplicaciones médicas, especialmente en el análisis de imágenes diagnósticas como radiografías, tomografías o resonancias magnéticas, los modelos de aprendizaje profundo (deep learning) procesan millones de parámetros internos que interactúan entre sí para identificar patrones visuales complejos y debido a esta enorme cantidad de variables y operaciones matemáticas, incluso los desarrolladores del sistema pueden tener dificultades para explicar exactamente cómo el algoritmo llegó a un diagnóstico específico (Topol, 2019). Esta falta de transparencia representa un desafío importante en el ámbito clínico, donde la interpretación médica tradicional se basa en procesos de razonamiento claros y justificables.

Figura 3

Caja negra y ataque adversarial en radiología.



Nota. La caja negra representa uno de los desafíos más importantes para los desarrolladores de IA en salud, ya que la precisión del algoritmo depende de la calidad de los patrones analizados. Fuente. Soffer V. et al. (2023).

Funcionamiento de la "Caja Negra" en el Análisis de Imágenes Médicas

Los sistemas de inteligencia artificial utilizados en radiología suelen emplear redes neuronales convolucionales, que son algoritmos diseñados para reconocer patrones visuales dentro de imágenes médicas. Estas redes analizan múltiples capas de información, desde características simples como bordes o contrastes hasta estructuras más complejas relacionadas con posibles patologías (Hosny et al., 2018).

El proceso general puede describirse en varias etapas

Ingreso de Datos: El sistema recibe imágenes médicas digitalizadas, como radiografías o tomografías.

Procesamiento Algorítmico: El algoritmo analiza las imágenes mediante múltiples capas neuronales que extraen patrones y características relevantes.

Generación del Resultado: El sistema produce una predicción diagnóstica, por ejemplo, la probabilidad de que exista una lesión pulmonar o un tumor.

Validación Clínica: El radiólogo revisa la predicción generada por el algoritmo y determina si coincide con su interpretación profesional.

El problema central radica en que, aunque el sistema pueda indicar que existe una alta probabilidad de enfermedad, muchas veces no es capaz de explicar claramente qué características específicas de la imagen condujeron a esa conclusión (Samek, Wiegand & Müller, 2017).

Implicaciones de la "Caja Negra" en la Seguridad del Paciente

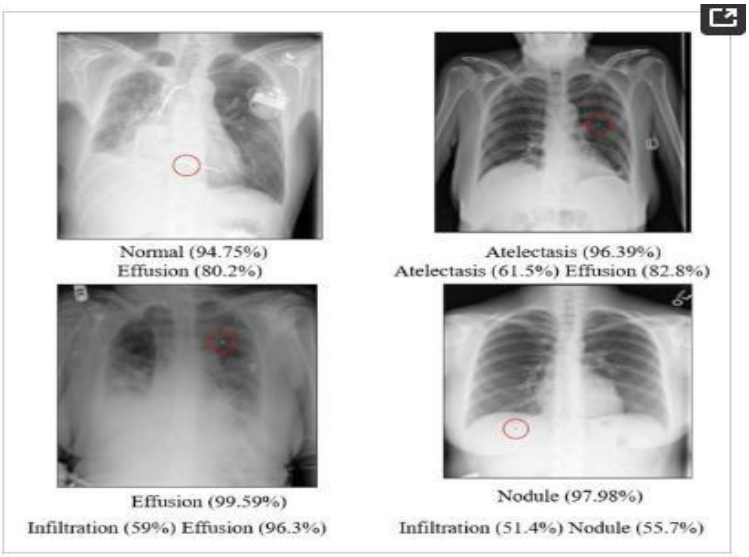
La opacidad de los sistemas de IA plantea múltiples desafíos en el ámbito médico, en primer lugar, dificulta que los profesionales de la salud comprendan el razonamiento detrás de una predicción algorítmica, lo que puede generar desconfianza en la tecnología o dependencia

excesiva de sus resultados, además, si el sistema comete un error diagnóstico, puede resultar complejo identificar el origen del fallo, y este problema se agrava cuando los algoritmos son entrenados con bases de datos limitadas o sesgadas, lo que puede generar resultados incorrectos al analizar imágenes provenientes de poblaciones diferentes (Finlayson et al., 2019).

Otro riesgo asociado es la posibilidad de ataques adversariales, en los cuales pequeñas modificaciones imperceptibles en las imágenes médicas pueden alterar significativamente la interpretación del algoritmo, donde estos ataques pueden provocar diagnósticos erróneos o manipulación de resultados clínicos, comprometiendo la seguridad del paciente y la integridad de los datos médicos (Finlayson et al., 2019).

Figura 4.

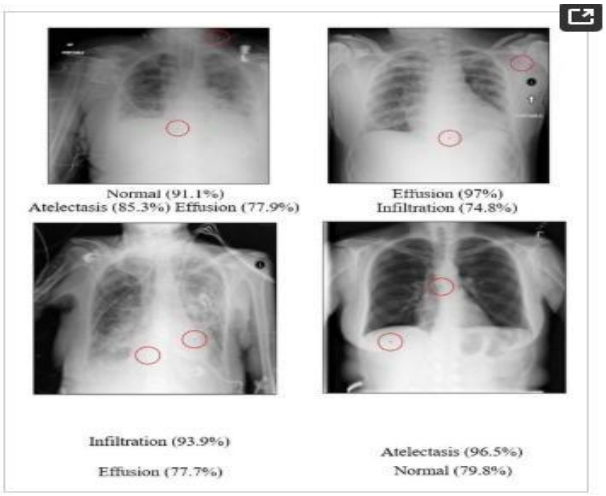
Ejemplos de ataques de un pixel al pecho exitoso.



Nota. Los círculos rojos indican los pixeles alterados durante los ataques de un pixel. Fuente. Mdpi.com. (2025)

Figura 5.

Ejemplos de ataques de dos pixeles al pecho exitosos.



Nota. Los círculos rojos indican los pixeles alterados durante los ataques de dos pixeles. Fuente Mdpi.com. (2025)

El Error Diagnóstico y la Seguridad del Paciente

El error diagnóstico continúa siendo uno de los principales desafíos en la seguridad del paciente y se entiende como la falta de establecer una explicación correcta y oportuna del problema de salud o de comunicarla adecuadamente. A partir de la evidencia reciente, se reconoce que este fenómeno no solo está asociado a limitaciones del razonamiento humano, sino también a factores sistémicos y tecnológicos. En este contexto, la incorporación de sistemas de inteligencia artificial en la práctica clínica ha ampliado la comprensión tradicional del error diagnóstico, al introducir nuevas fuentes de riesgo relacionadas con la calidad de los datos, el diseño de los algoritmos y su comportamiento en entornos reales. Analizando la literatura actual, se evidencia que la seguridad del paciente depende ahora de la interacción entre el juicio clínico y el desempeño de los sistemas automatizados, lo que implica considerar variables como la robustez del modelo, su capacidad de generalización y la existencia de mecanismos de

supervisión y validación continua. En consecuencia, el error diagnóstico en entornos asistidos por inteligencia artificial debe entenderse como el resultado de una interacción compleja entre factores humanos, tecnológicos y organizacionales, lo cual exige enfoques integrales para su prevención (Singh et al., 2019; WHO, 2021; Rajpurkar et al., 2022).

Taxonomía del Error: Error Humano vs. Error Algorítmico

La literatura actual sobre razonamiento clínico ha demostrado que el error humano en medicina se asocia con frecuencia a la presencia de sesgos cognitivos que surgen del funcionamiento dual del pensamiento. Analizando la evidencia reciente, se reconoce que los profesionales de la salud alternan entre un sistema intuitivo, rápido y basado en la experiencia y el reconocimiento de patrones, y un sistema analítico, más lento y reflexivo, orientado a la evaluación crítica de la información. Si bien este modelo permite una toma de decisiones eficiente en contextos clínicos complejos, también puede favorecer errores cuando predominan respuestas automáticas o cuando existen limitaciones en el conocimiento o en la interpretación de los datos disponibles. En este sentido, los estudios contemporáneos resaltan que la interacción entre ambos sistemas, junto con factores contextuales, influye directamente en la precisión diagnóstica (Norman et al., 2021). Entre los sesgos más comunes se encuentran el anclaje, el cierre prematuro y la confirmación selectiva de hipótesis. No obstante, en el contexto de la IA autónoma emerge una categoría distinta de error, como el error algorítmico. A diferencia del error humano, que surge de procesos cognitivos individuales, el error algorítmico tiene origen estructural y sistémico, y puede derivarse de datos de entrenamiento no representativos, sesgos incorporados en los conjuntos de datos, sobreajuste del modelo o incapacidad para generalizar adecuadamente a poblaciones distintas de aquellas utilizadas en su validación (Obermeyer et al., 2019; Kelly et al., 2019). Este tipo de error no responde a fatiga, intuición o juicio clínico

deficiente, sino a limitaciones inherentes al diseño y entrenamiento del sistema. La distinción entre error humano y error algorítmico no implica que ambos operen de manera aislada, sino que por el contrario, en entornos clínicos híbridos, el error puede surgir de la interacción entre ambos sistemas, donde un profesional puede confiar excesivamente en una predicción automatizada sin cuestionarla críticamente, o puede desestimar una recomendación válida por desconfianza injustificada, y en este sentido, el error diagnóstico contemporáneo adquiere una naturaleza sociotécnica, donde la responsabilidad y el análisis causal deben considerar tanto la dimensión humana como la tecnológica (Benjamens et al., 2020).

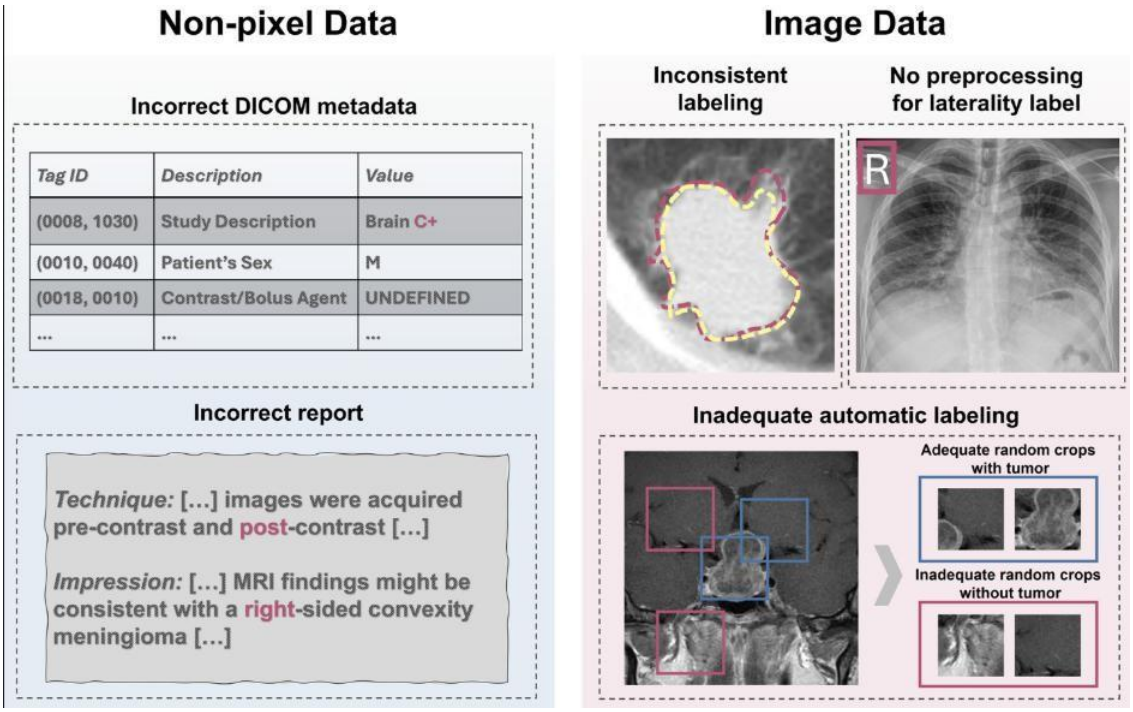
Sesgos de Automatización en la Práctica Clínica

La integración de sistemas de IA en el proceso diagnóstico ha puesto de relieve el fenómeno conocido como sesgo de automatización (automation bias), ya que este sesgo se manifiesta cuando el profesional medico tiende a aceptar de manera acrítica la recomendación de un sistema automatizado, incluso en presencia de información contradictoria o dudas razonables (Goddard et al., 2012). En otras palabras, la confianza en la precisión tecnológica puede disminuir el nivel de vigilancia cognitiva del profesional, y, por otra parte, también puede presentarse el fenómeno opuesto: el rechazo sistemático de la recomendación algorítmica, aun cuando esta sea correcta. Este comportamiento puede estar asociado a desconfianza hacia la tecnología, desconocimiento de su funcionamiento o temor a la pérdida de control profesional, y en ambos extremos la sobreconfianza y el rechazo injustificado pueden comprometer la calidad diagnóstica. Estudios empíricos han demostrado que la interacción entre humanos y sistemas automatizados modifica la dinámica del juicio clínico, pudiendo inducir dependencia cognitiva o delegación excesiva de responsabilidad (Cabitza et al., 2017). En consecuencia, la seguridad del paciente no depende únicamente de la precisión estadística del modelo, sino también del diseño

de interfaces, la formación del personal sanitario y la existencia de protocolos que promuevan una supervisión crítica. Desde una perspectiva ética, la presencia de sesgos de automatización plantea interrogantes sobre el estándar de diligencia exigible al profesional que utiliza IA. Si bien la tecnología puede mejorar la capacidad diagnóstica, no exime al médico de su deber de evaluación independiente y razonada, y la literatura sugiere que la IA debe concebirse como un sistema colaborativo que complemente, pero no sustituya completamente, el juicio clínico humano (Topol, 2019).

Figura 6

Ejemplos de ataques adversariales que dan un reporte incorrecto.



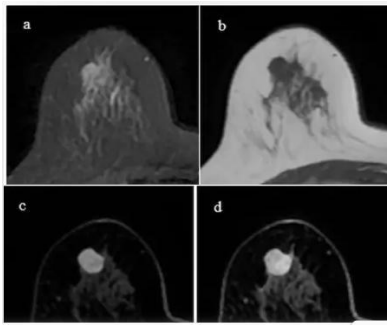
Nota. Una de las consecuencias más graves de los ataques adversariales corresponde a las fallas en los reportes médicos, lo que puede afectar de manera sustancial la seguridad de los pacientes. Fuente. Vasan, D. & Hammoudeh, M. (2024).

Impacto Clínico: Falsos Positivos y Falsos Negativos en Sistemas Autónomos

El desempeño diagnóstico de los sistemas de IA suele evaluarse mediante métricas como sensibilidad, especificidad y valores predictivos, sin embargo, más allá de los indicadores estadísticos, resulta fundamental analizar las consecuencias clínicas de los errores, particularmente en términos de falsos positivos y falsos negativos. Un falso positivo ocurre cuando el sistema identifica erróneamente la presencia de una enfermedad, y en el contexto clínico, esto puede generar pruebas adicionales innecesarias, tratamientos injustificados, ansiedad en el paciente y aumento de costos sanitarios. En especialidades como la oncología o la radiología, un falso positivo puede conducir a procedimientos invasivos con riesgos asociados. Por su parte, un falso negativo implica la omisión de una patología existente, y este tipo de error suele tener consecuencias más graves, ya que puede retrasar el tratamiento oportuno y agravar el pronóstico del paciente. En sistemas autónomos, el riesgo se amplifica cuando la decisión algorítmica no es revisada de manera inmediata por un profesional, lo que puede permitir que el error se materialice sin corrección temprana (He et al., 2019). Además, cuando un modelo presenta sesgos sistemáticos derivados de datos no representativos, los falsos negativos o positivos pueden afectar de manera desproporcionada a determinados grupos poblacionales, generando inequidades en la atención sanitaria (Obermeyer et al., 2019). En este sentido, el impacto clínico del error algorítmico no se limita al individuo, sino que puede reproducir desigualdades estructurales en el sistema de salud. En consecuencia, la evaluación del riesgo en sistemas autónomos no debe centrarse exclusivamente en la precisión global del modelo, sino en el análisis contextualizado de sus consecuencias clínicas, la gravedad potencial del daño y la posibilidad de mitigación mediante supervisión humana significativa.

Figura 7

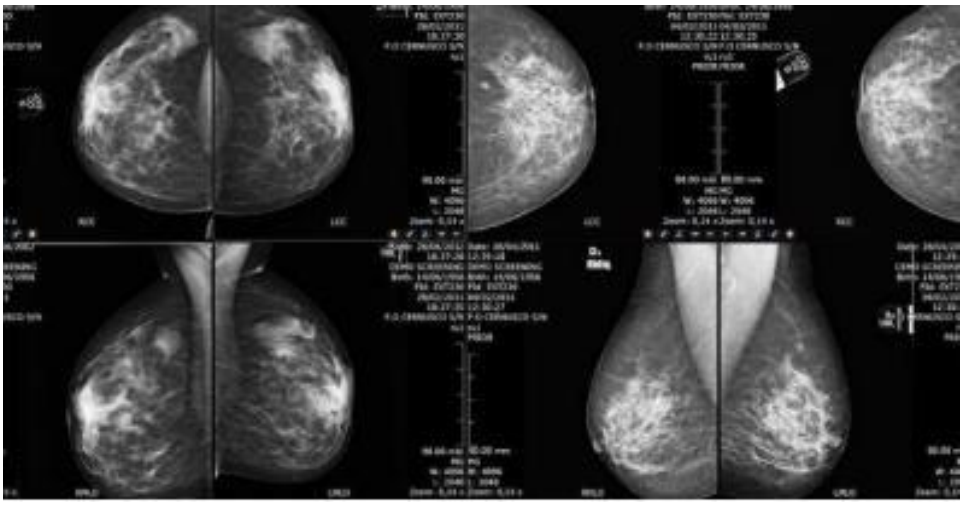
Falsos negativos y positivos en la Resonancia Magnética de Mama



Nota. La lesión de la masa mamaria derecha fue interpretada originalmente como benigna, por sus características homogéneas. Al realizar el estudio histopatológico, se evidenció que la masa era maligna. Fuente. Latam, R. (2024).

Figura 8

Falsos negativos y positivos en la Resonancia Magnética de Mama



Nota. El uso de patrones en IA debe ser estudiado y auditado a profundidad, para evitar errores ocasionados por falsos positivos o falsos negativos. Fuente Eddy D.M (1982).

A modo de ejemplo, se presenta el siguiente caso: una mujer que va a su médico, o médica, de cabecera porque nota que tiene un bulto en uno de sus pechos, ante la sospecha

clinica, el facultativo se solicita la realización de una mamografía y finalmente esta es positiva (efectivamente se detecta una masa sospechosa). De nuevo, se plantea determinar la probabilidad de que la mujer tenga cáncer realmente, o de que sea un “falso positivo”.

La tabla siguiente muestra la fiabilidad de las mamografías en el diagnóstico de masas benignas y malignas. Además, la probabilidad de que esa masa sea maligna, es decir, de que la mujer tenga un cáncer de mama, se estimó en el artículo que era de un 1%. La incidencia es de 1 de cada 100 mujeres.

Tabla 2

Fiabilidad de la mamografía en la detección de masas malignas y benignas

Mamografía	Masa maligna	Masa benigna
Positiva	79.2%	9.6%
Negativa	20.8%	90.4%

Nota. La mamografía tiene una alta sensibilidad para la detección de masas malignas. Es importante estudiar las ventajas del uso de la IA en este contexto. Fuente.Eddy D.M (1982)

Para analizar la probabilidad de falsos positivos, se va a suponer que 10.000 mujeres con un bulto en el pecho se realizan una mamografía. Como la incidencia del cáncer de mama es del 1%, entonces habrá unas 100 que tengan cáncer y 9.900 que no. Al realizar la mamografía se evidenciará:

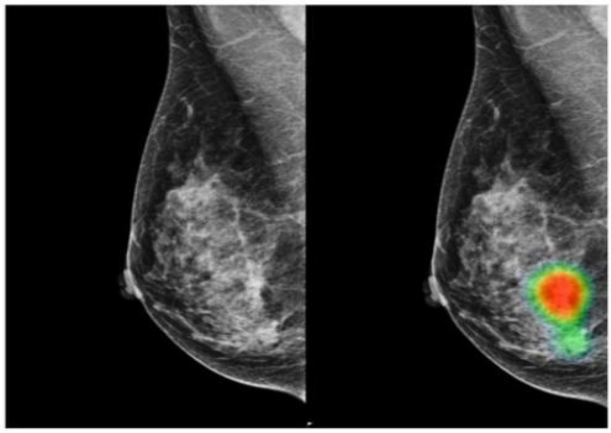
De las 100 mujeres que tienen cáncer de mama, 79 obtendrán un resultado positivo, es decir, la masa en el pecho es sospechosa de malignidad (cáncer), mientras que para el resto, 21 mujeres, la prueba indicará que corresponde a una masa benigna.

De las 9.900 mujeres que realmente no tienen cáncer de mama, la mamografía revelará que es una masa sospechosa de malignidad para 950 mujeres y el resto, 8.950, se clasificarán como benignas.

De las 10.000 mamografías realizadas, 950 + 79 = 1.029 han reportado un resultado positivo, la masa sospechosa parece ser maligna según la mamografía, de las cuales 79 corresponden realmente a una masa maligna, a un cáncer de mama, luego la probabilidad de que una mujer tenga realmente cáncer de mama si la mamografía le ha dado positiva es de 79/1029, es decir, alrededor de un 7,7%. O equivalentemente, las mamografías tienen un porcentaje del 92,3% de falsos positivos (Tomé, C. 2015).

Figura 9

Reconocimiento de patrones tumorales en mamografía mediante IA



Nota. Mamografía de paciente de 49 años con carcinma lobulillar de mama derecha. La masa en color naranja fue detectada por IA con una puntuación de anormalidad de 96%. Fuente. Lunit. (2020).

Ética Médica y Bioética frente a la Autonomía de las Máquinas

La progresiva incorporación de sistemas autónomos de inteligencia artificial en el ámbito clínico no solo plantea desafíos técnicos y regulatorios, sino que interpela directamente los

fundamentos de la ética médica y la bioética contemporánea, donde, la posibilidad de que una máquina participe activamente en decisiones diagnósticas o terapéuticas obliga a revisar categorías tradicionales como agencia moral, responsabilidad profesional y deber de cuidado, y en la práctica médica clásica, la toma de decisiones está intrínsecamente vinculada a la deliberación ética del profesional, quien asume responsabilidad moral y jurídica por sus actos, pero, sin embargo, cuando intervienen sistemas autónomos capaces de generar decisiones sin intervención humana directa, emerge una tensión entre autonomía tecnológica y control humano, y esta tensión ha sido objeto de análisis en la literatura reciente, que advierte que la innovación tecnológica no puede desvincularse de los principios éticos fundamentales que rigen la práctica sanitaria (Topol, 2019; Morley et al., 2020). Donde la cuestión central no es únicamente si la IA puede mejorar la precisión diagnóstica, sino si su incorporación respeta los valores que históricamente han orientado la medicina: protección del paciente, justicia distributiva y responsabilidad profesional.

Principialismo Bioético en la Era Digital

El principialismo bioético, formulado por Beauchamp y Childress, ha constituido uno de los marcos normativos más influyentes en la ética biomédica contemporánea. Sus cuatro principios beneficencia, no maleficencia, autonomía y justicia adquieren nuevas dimensiones en el contexto de la medicina digital (Beauchamp & Childress, 2019).

Beneficencia: Este principio exige que las intervenciones médicas promuevan el bienestar del paciente y en el contexto de la IA, este principio se traduce en la obligación de implementar sistemas cuya eficacia esté clínicamente validada y cuyo uso incremente la calidad diagnóstica o terapéutica, pero, no obstante, la mera superioridad estadística del algoritmo no es

suficiente; debe evaluarse también su impacto real en resultados clínicos y su integración responsable en el entorno asistencial (Topol, 2019).

No maleficencia: Este principio implica evitar daños previsibles, en sistemas autónomos, el riesgo puede derivarse de sesgos algorítmicos, errores de generalización o falta de supervisión adecuada (Obermeyer et al., 2019). La opacidad de la "caja negra" intensifica este desafío, pues dificulta la identificación temprana de fallas estructurales (Rudin, 2019). En consecuencia, el deber de no causar daño exige mecanismos robustos de validación, auditoría continua y trazabilidad.

Autonomía: En el paciente se vincula con el derecho a recibir información comprensible y a tomar decisiones libres e informadas sobre su atención médica, la utilización de sistemas cuya lógica interna no puede explicarse plenamente plantea interrogantes sobre la calidad del consentimiento informado, si el profesional no puede justificar de manera clara el fundamento de una decisión generada por IA, la capacidad del paciente para comprender y aceptar el tratamiento puede verse comprometida (Morley et al., 2020).

Justicia: Este principio adquiere especial relevancia frente a la evidencia de que los modelos algorítmicos pueden reproducir desigualdades existentes si son entrenados con datos no representativos (Obermeyer et al., 2019). La implementación de IA en salud debe garantizar equidad en el acceso y evitar discriminaciones indirectas derivadas de sesgos estructurales, y en este sentido, la justicia en la era digital exige no solo eficiencia tecnológica, sino también responsabilidad social en el diseño y despliegue de los sistemas.

El Concepto de "Human-in-the-loop": Supervisión Humana como Imperativo Ético

El modelo human-in-the-loop se ha consolidado como un principio rector en la gobernanza ética de la inteligencia artificial aplicada a la salud. Este enfoque sostiene que,

incluso cuando se utilicen sistemas altamente automatizados, debe mantenerse una intervención humana significativa en la toma de decisiones (European Commission, 2021). Desde el punto de vista ético, la supervisión humana no constituye únicamente un requisito técnico, sino un imperativo moral, la medicina es una práctica intrínsecamente relacional, basada en la confianza entre paciente y profesional, donde delegar completamente decisiones clínicas en una máquina podría debilitar esta relación y erosionar la responsabilidad moral del médico.

Además, la presencia de un profesional en el circuito decisorio permite contextualizar la recomendación algorítmica, integrar factores subjetivos del paciente y ejercer juicio prudencial, la literatura enfatiza que la IA debe concebirse como herramienta de apoyo colaborativo y no como sustituto absoluto del juicio clínico (Topol, 2019). En consecuencia, la supervisión humana significativa opera como mecanismo de contención frente al riesgo de automatización acrítica y como garantía de rendición de cuentas.

Responsabilidad Moral vs. Responsabilidad Técnica: ¿Puede una Máquina Ser Sujeto de Juicio Ético?

Uno de los debates más complejos en la ética de la IA médica gira en torno a la posibilidad de atribuir responsabilidad moral a un sistema autónomo, y desde la filosofía moral clásica, la responsabilidad implica intencionalidad, conciencia y capacidad de deliberación, los sistemas de IA, aunque puedan simular procesos de toma de decisiones, carecen de conciencia y de comprensión moral genuina (Rudin, 2019) y por ello, la mayoría de la doctrina sostiene que la IA no puede ser considerada sujeto moral, sino instrumento técnico cuya responsabilidad recae en los agentes humanos involucrados en su diseño, implementación y utilización (Morley et al., 2020). Esta distinción entre responsabilidad moral y responsabilidad técnica resulta fundamental: mientras la máquina ejecuta operaciones basadas en datos y algoritmos, la evaluación ética y la

asunción de consecuencias corresponden a personas físicas o jurídicas, y no obstante, la creciente autonomía funcional de los sistemas plantea desafíos prácticos en la atribución de responsabilidades, especialmente cuando el error resulta de interacciones complejas entre desarrolladores, instituciones sanitarias y profesionales clínicos. En este contexto, la ética médica contemporánea debe articular mecanismos que aseguren trazabilidad, transparencia y claridad en la distribución de deberes.

En definitiva, aunque la IA pueda participar activamente en el proceso diagnóstico, la dimensión moral del acto médico continúa siendo indelegable. La tecnología puede ampliar las capacidades humanas, pero no sustituye la responsabilidad ética inherente al ejercicio profesional.

Marco Regulatorio Internacional de la Inteligencia Artificial en Salud

La consolidación de sistemas autónomos de inteligencia artificial (IA) en el ámbito sanitario ha impulsado la creación de marcos regulatorios internacionales orientados a garantizar la seguridad del paciente, la transparencia tecnológica y la rendición de cuentas, dado que la IA médica puede influir directamente en decisiones diagnósticas y terapéuticas, su regulación no puede limitarse a estándares técnicos generales, sino que debe integrarse dentro del régimen jurídico propio de los dispositivos médicos y de la protección de derechos fundamentales. A nivel internacional, destacan tres referentes normativos y de gobernanza: el Reglamento de Inteligencia Artificial de la Unión Europea, las directrices de la Organización Mundial de la Salud (OMS) y el enfoque regulatorio de la Food and Drug Administration (FDA) de los Estados Unidos en relación con el Software as a Medical Device (SaMD), y con cada uno de estos instrumentos aborda la IA sanitaria desde perspectivas complementarias, combinando exigencias técnicas, principios éticos y mecanismos de supervisión continua.

Reglamento de Inteligencia Artificial de la Unión Europea

La Unión Europea ha desarrollado uno de los marcos regulatorios más avanzados en materia de inteligencia artificial mediante el denominado Artificial Intelligence Act, y este reglamento adopta un enfoque basado en el riesgo, clasificando los sistemas de IA según su potencial impacto sobre derechos fundamentales y seguridad (European Commission, 2021), y en este esquema, los sistemas de IA utilizados en el diagnóstico médico o que influyen en decisiones relacionadas con la salud son categorizados como sistemas de "alto riesgo", donde esta clasificación implica la imposición de obligaciones estrictas antes de su comercialización y durante todo su ciclo de vida, en los cuales, dentro de los requisitos principales se encuentran, sistemas robustos de gestión de riesgos, garantías sobre la calidad y gobernanza de los datos utilizados en el entrenamiento, documentación técnica detallada que permita trazabilidad, transparencia en el funcionamiento del sistema, supervisión humana significativa, vigilancia postcomercialización continua. La inclusión de la IA médica en la categoría de alto riesgo refleja el reconocimiento de que los errores algorítmicos pueden afectar directamente la vida y la integridad de las personas y asimismo, el reglamento vincula la regulación tecnológica con la protección de derechos fundamentales, como la igualdad y la no discriminación, especialmente frente al riesgo de sesgos estructurales en los datos (European Commission, 2021).

Desde una perspectiva bioética, este enfoque normativo refuerza el principio de no maleficencia y la obligación institucional de anticipar riesgos antes de que se materialicen en daños clínicos. No se trata únicamente de reaccionar ante errores, sino de prevenirlos mediante controles ex ante y mecanismos de supervisión estructurados.

Directrices de la Organización Mundial de la Salud (OMS)

La Organización Mundial de la Salud (OMS) ha abordado la inteligencia artificial en salud desde una perspectiva centrada en la ética y la gobernanza global, en su informe sobre ética y gobernanza de la IA para la salud, la OMS establece principios orientadores que buscan garantizar que el desarrollo y la implementación de estas tecnologías respeten los derechos humanos y promuevan la equidad sanitaria (World Health Organization [WHO], 2021). Entre los principios principales se encuentran inicialmente protección de la autonomía humana, promoción del bienestar y la seguridad pública, garantía de transparencia y explicabilidad, fomento de la responsabilidad y la rendición de cuentas, aseguramiento de la equidad e inclusión, promoción de sistemas sostenibles y adaptables.

Y a diferencia del enfoque estrictamente jurídico de la Unión Europea, las directrices de la OMS tienen un carácter orientador y ético, proporcionando un marco para que los Estados diseñen sus propias políticas regulatorias, y, sin embargo, su relevancia radica en que enfatizan la necesidad de integrar la gobernanza tecnológica con valores fundamentales de la salud pública, especialmente en contextos de desigualdad estructural. La OMS subraya que la implementación de IA no debe ampliar brechas existentes en acceso a la atención sanitaria ni generar nuevas formas de exclusión y, asimismo, insiste en la necesidad de participación interdisciplinaria en el diseño y supervisión de estos sistemas, reconociendo que la gobernanza de la IA en salud requiere la colaboración entre profesionales sanitarios, ingenieros, juristas y expertos en ética.

Estándares de la FDA (Estados Unidos): Enfoque de Ciclo de Vida del SaMD

En los Estados Unidos, la regulación de la IA médica se ha desarrollado principalmente a través de la clasificación de estos sistemas como Software as a Medical Device (SaMD). La FDA

ha adoptado un enfoque centrado en la evaluación de seguridad y eficacia, exigiendo evidencia clínica antes de la comercialización (FDA, 2019). Un elemento distintivo del modelo estadounidense es su énfasis en el ciclo de vida total del producto (*total product lifecycle approach*). Este enfoque reconoce que los sistemas basados en aprendizaje automático pueden evolucionar con el tiempo, especialmente aquellos que incorporan mecanismos de actualización continua, por ello, la regulación no se limita a la fase previa a la aprobación, sino que incluye monitoreo postcomercialización, vigilancia de desempeño y control de modificaciones algorítmicas.

La FDA ha propuesto además el denominado Predetermined Change Control Plan, un mecanismo que permite anticipar y regular modificaciones futuras del algoritmo sin requerir necesariamente una nueva aprobación completa, siempre que dichas modificaciones se encuentren previamente delimitadas y justificadas (FDA, 2019). Este modelo refleja el reconocimiento de que la autonomía algorítmica no es estática, sino evolutiva, y desde la perspectiva de la seguridad del paciente, el enfoque del ciclo de vida resulta particularmente relevante, pues admite que el error diagnóstico puede emerger incluso después de la validación inicial del sistema, y en consecuencia, la regulación debe ser dinámica y adaptativa, en consonancia con la naturaleza cambiante de los modelos de aprendizaje automático.

Marco Regulatorio Nacional

La implementación de sistemas autónomos de inteligencia artificial (IA) en el ámbito médico no solo debe analizarse desde la perspectiva internacional, sino también a la luz del ordenamiento jurídico nacional, la protección del paciente frente a fallos tecnológicos, el tratamiento de datos clínicos utilizados para entrenar algoritmos y la certificación de dispositivos médicos constituyen ejes fundamentales de regulación interna, en el contexto colombiano, el

marco normativo no contempla aún una ley específica sobre inteligencia artificial en salud; sin embargo, diversas disposiciones vigentes permiten articular un régimen aplicable a estos sistemas a partir de normas sobre derechos del paciente, protección de datos personales y regulación de dispositivos médicos, esta aproximación normativa indirecta obliga a interpretar el uso de la IA dentro de estructuras jurídicas preexistentes.

Leyes de Derechos y Deberes de los Pacientes: Protección frente a Fallos Tecnológicos

El reconocimiento de los derechos del paciente constituye el fundamento jurídico de cualquier análisis sobre responsabilidad en salud, en Colombia, la Ley Estatutaria 1751 de 2015 consagra el derecho fundamental a la salud, estableciendo que los servicios sanitarios deben prestarse bajo criterios de calidad, seguridad, oportunidad y continuidad (Congreso de la República de Colombia, 2015). Estos principios resultan plenamente aplicables cuando intervienen sistemas tecnológicos en el proceso asistencial. Asimismo, la Carta de Derechos y Deberes del Paciente, adoptada mediante la Resolución 229 de 2020 del Ministerio de Salud y Protección Social, reconoce el derecho del usuario a recibir información clara, comprensible y suficiente sobre los procedimientos diagnósticos y terapéuticos (Ministerio de Salud y Protección Social, 2020). En el contexto de la IA, este derecho se vincula directamente con la obligación de informar sobre la utilización de sistemas automatizados en el diagnóstico y sus posibles riesgos.

La protección frente a fallos tecnológicos se deriva del principio de seguridad del paciente, además, si un sistema autónomo genera un error diagnóstico, la institución prestadora del servicio de salud mantiene la obligación de garantizar atención segura y de responder por fallas en la prestación del servicio, y en este sentido, la incorporación de IA no exime a los actores sanitarios de su deber de diligencia, sino que amplía el estándar de cuidado, exigiendo verificación, supervisión y evaluación continua de las herramientas tecnológicas empleadas.

Normativa de Protección de Datos Personales: Uso de Datos Clínicos para el Entrenamiento de la IA

El desarrollo y entrenamiento de sistemas de IA médica requieren grandes volúmenes de datos clínicos, incluyendo historias médicas, imágenes diagnósticas y resultados de laboratorio, en este tratamiento masivo de información plantea desafíos significativos en materia de privacidad y protección de datos personales. En Colombia, la Ley 1581 de 2012 establece el régimen general de protección de datos personales, reconociendo el derecho fundamental al habeas data y regulando el tratamiento de información sensible, categoría dentro de la cual se incluyen los datos relativos a la salud (Congreso de la República de Colombia, 2012), lo que la norma exige consentimiento previo, informado y expreso del titular, salvo excepciones previstas por la ley.

El carácter sensible de los datos clínicos implica mayores exigencias de seguridad, confidencialidad y finalidad específica en su uso, y en el contexto del entrenamiento algorítmico, surge la necesidad de garantizar que la información sea anonimizada o pseudonimizada cuando sea posible, reduciendo el riesgo de identificación indebida. Además, el principio de finalidad obliga a que los datos sean utilizados exclusivamente para los propósitos autorizados, lo que plantea interrogantes sobre el uso secundario de información clínica con fines de desarrollo tecnológico, y, desde una perspectiva bioética, la protección de datos se relaciona con el principio de autonomía y con el respeto por la dignidad del paciente. Asimismo, la explotación masiva de datos sin garantías adecuadas puede erosionar la confianza en el sistema sanitario y comprometer la legitimidad de la innovación tecnológica.

Regulación de Dispositivos Médicos: Certificación y Vigilancia Post-Mercado

La inteligencia artificial aplicada al diagnóstico médico puede clasificarse jurídicamente como software con finalidad sanitaria, lo que la ubica dentro del régimen de dispositivos médicos, y en Colombia, el Decreto 4725 de 2005 regula el régimen de registros sanitarios, vigilancia y control de dispositivos médicos, estableciendo la obligación de certificación previa a su comercialización (Ministerio de la Protección Social, 2005). El Instituto Nacional de Vigilancia de Medicamentos y Alimentos (INVIMA) es la autoridad competente para evaluar la seguridad, eficacia y calidad de estos productos antes de su autorización, en el caso de sistemas basados en IA, la evaluación debe considerar no solo la funcionalidad técnica, sino también la evidencia clínica que respalde su desempeño diagnóstico.

Un elemento clave del régimen nacional es la vigilancia post-mercado, una vez autorizado el dispositivo, el fabricante y las instituciones usuarias tienen la obligación de reportar eventos adversos y fallas técnicas, este mecanismo resulta especialmente relevante para sistemas con componentes de aprendizaje automático, cuya eficacia puede variar según el contexto clínico o la población en la que se implementen. La vigilancia continua permite detectar errores sistemáticos, corregir fallas y retirar del mercado tecnologías que representen riesgos para la seguridad del paciente. En consecuencia, el régimen de dispositivos médicos constituye un instrumento esencial para mitigar los efectos de errores algorítmicos y garantizar la protección del usuario frente a innovaciones tecnológicas.

La Responsabilidad Jurídica por Error de Inteligencia Artificial en el Diagnóstico Médico

La incorporación de sistemas autónomos de inteligencia artificial (IA) en el ámbito sanitario ha generado una reconfiguración del análisis tradicional de la responsabilidad jurídica, y cuando un error diagnóstico produce un daño al paciente, surge el interrogante sobre quién

debe asumir las consecuencias jurídicas del perjuicio A diferencia del modelo clásico, donde el acto médico se atribuía exclusivamente al profesional, la intervención de un algoritmo introduce una dimensión tecnológica que complejiza la imputación. El problema no radica únicamente en determinar la existencia del daño, sino en identificar el sujeto responsable dentro de una red de actores que incluye al médico tratante, la institución prestadora de servicios de salud, el desarrollador del software y el fabricante del sistema, y la literatura ha señalado que los sistemas autónomos generan escenarios de "responsabilidad distribuida", en los cuales el resultado clínico es producto de una interacción sociotécnica (Price et al., 2019; Morley et al., 2020).

Desde el punto de vista jurídico, la responsabilidad puede analizarse principalmente desde dos grandes teorías: la responsabilidad por producto defectuoso y la responsabilidad por negligencia profesional (malpractice). Ambas categorías ofrecen marcos interpretativos distintos para abordar el daño derivado del error algorítmico.

Responsabilidad Civil: ¿Quién Responde por el Daño?

La responsabilidad civil en el ámbito sanitario tiene como finalidad reparar el daño causado al paciente cuando se acredita una conducta antijurídica, un perjuicio y un nexo causal entre ambos, en el contexto de la IA médica, la determinación del responsable depende del origen del error y del grado de intervención humana en la decisión final. Si el sistema autónomo actúa como herramienta de apoyo bajo supervisión activa del médico (human-in-the-loop), es probable que la responsabilidad recaiga principalmente en el profesional o en la institución sanitaria, especialmente si se demuestra falta de diligencia en la interpretación o validación del resultado, en cambio, cuando el sistema opera con autonomía significativa (human-on-the-loop), el análisis puede extenderse hacia el desarrollador o fabricante del software, en la medida en que el error derive de defectos estructurales del algoritmo.

La doctrina contemporánea advierte que la simple utilización de tecnología avanzada no exonera automáticamente al profesional sanitario de su deber de cuidado (Topol, 2019)., no obstante, tampoco resulta jurídicamente razonable atribuirle responsabilidad por fallas técnicas invisibles o imprevisibles derivadas de la arquitectura interna del modelo, y en consecuencia, el análisis debe considerar la previsibilidad del riesgo, el cumplimiento de protocolos de validación y el estándar profesional exigible en cada caso.

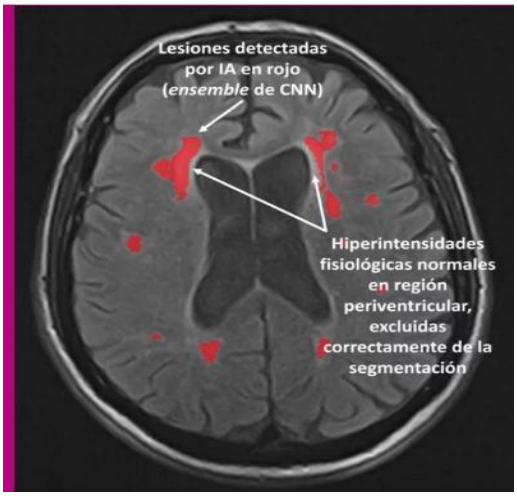
Responsabilidad por Producto Defectuoso: La IA como Objeto

La teoría de la responsabilidad por producto defectuoso parte de la premisa de que el fabricante o proveedor debe responder por los daños causados por un producto que no ofrece la seguridad que legítimamente se espera de él, y ha sido tradicionalmente aplicada a bienes materiales, esta doctrina ha sido progresivamente extendida al software con finalidad médica (European Commission, 2021). Bajo esta perspectiva, la IA diagnóstica puede considerarse un producto, si el algoritmo presenta defectos de diseño, errores en el entrenamiento o fallas sistemáticas que comprometen su seguridad, el desarrollador o fabricante podría ser responsable por los daños ocasionados. Este enfoque resulta particularmente relevante cuando el error se origina en sesgos de datos, sobreajuste o fallas de generalización no advertidas adecuadamente (Obermeyer et al., 2019; Kelly et al., 2019). Sin embargo, la aplicación de esta teoría enfrenta desafíos específicos, a diferencia de un dispositivo físico, los sistemas de IA pueden modificar su comportamiento mediante actualizaciones o aprendizaje continuo, lo que plantea interrogantes sobre el momento en que debe evaluarse el defecto y sobre la distribución de responsabilidades entre desarrollador, proveedor de datos y entidad que implementa el sistema, además, la opacidad algorítmica puede dificultar la identificación precisa del defecto. Si el modelo opera

como “caja negra”, demostrar que el daño se debió a una falla estructural del producto puede requerir peritajes técnicos complejos (Rudin, 2019).

Figura 10

Detección de lesiones de sustancia blanca cerebral con algoritmo basado en ensamble CNN.



Nota. Detección de lesiones de la sustancia blanca utilizando algoritmos CNN. Fuente. Valverde, S. et al. (2017)

Responsabilidad por Negligencia Profesional (Malpractice): El Médico como Usuario

La responsabilidad por negligencia profesional se configura cuando el médico incumple el estándar de cuidado exigible en su práctica, causando daño al paciente. En el contexto de la IA, este estándar debe reinterpretarse considerando la disponibilidad y uso de herramientas tecnológicas avanzadas, si el profesional acepta de manera acrítica una recomendación algorítmica sin ejercer juicio clínico independiente, podría configurarse una falla por exceso de confianza o por delegación indebida de la decisión (Goddard et al., 2012). Por el contrario, si el médico ignora sistemáticamente una herramienta validada y ello conduce a un daño evitable, también podría discutirse una omisión negligente.

La cuestión central consiste en determinar cuál es el estándar profesional razonable en entornos clínicos digitalizados, donde la literatura sugiere que la diligencia médica en la era de la IA implica no solo conocimientos clínicos, sino también comprensión básica del funcionamiento y limitaciones de los sistemas utilizados (Benjamens et al., 2020). Así, la formación tecnológica se convierte en un componente del deber de cuidado, no obstante, la responsabilidad por malpractice no puede extenderse ilimitadamente al profesional cuando el error proviene de defectos técnicos imposibles de detectar mediante supervisión razonable, donde la delimitación entre negligencia médica y falla tecnológica constituye uno de los principales desafíos interpretativos del derecho contemporáneo.

El Problema de la Causalidad: Dificultades Probatorias en Algoritmos Opacos

La determinación de responsabilidad civil exige acreditar el nexo causal entre la conducta imputada y el daño sufrido, y en sistemas de IA opacos, esta exigencia adquiere complejidad particular. La "caja negra" dificulta explicar por qué el algoritmo generó una predicción específica, lo que complica demostrar que dicha predicción fue la causa directa del perjuicio (Rudin, 2019). La dificultad probatoria se intensifica cuando el resultado clínico deriva de múltiples factores concurrentes: juicio médico, datos de entrada, configuración institucional y desempeño del modelo, en estos escenarios, la causalidad no es lineal, sino distribuida. Price et al. (2019) sostienen que los sistemas autónomos generan cadenas causales fragmentadas, donde ningún actor individual controla completamente el resultado.

Ante esta complejidad, algunos autores plantean la necesidad de flexibilizar los estándares probatorios o de aplicar criterios como la inversión de la carga de la prueba en determinados supuestos, especialmente cuando el paciente se encuentra en desventaja técnica para demostrar el defecto, y, asimismo, la exigencia de trazabilidad y registro de decisiones

algorítmicas impulsada por marcos regulatorios internacionales puede facilitar la reconstrucción causal y fortalecer la transparencia.

En definitiva, la causalidad en el contexto de IA médica requiere una reinterpretación que tenga en cuenta la naturaleza sociotécnica del acto diagnóstico y la atribución de responsabilidad no puede basarse únicamente en esquemas clásicos diseñados para relaciones bilaterales simples, sino que debe adaptarse a escenarios donde la decisión emerge de la interacción entre humanos y sistemas automatizados.

Análisis Comparativo: Ética vs. Norma en la Regulación de la Inteligencia Artificial Médica

La integración de sistemas de inteligencia artificial en el acto médico ha evidenciado una tensión estructural entre el plano ético y el plano normativo, y, mientras la ética médica se fundamenta en principios orientadores de carácter universal y valorativo, la norma jurídica opera mediante reglas concretas, coercibles y delimitadas territorialmente, ya que esta diferencia ontológica implica que no toda exigencia ética encuentra una traducción inmediata en disposiciones legales. La ética, particularmente desde el principialismo bioético, propone estándares amplios que orientan la conducta profesional incluso en ausencia de regulación específica (Beauchamp & Childress, 2019). Por el contrario, el derecho positivo requiere tipificación, delimitación de competencias y mecanismos probatorios para su aplicación, ya que en el contexto de la IA médica, esta divergencia se hace especialmente visible cuando la tecnología evoluciona con mayor rapidez que los marcos regulatorios.

El análisis comparativo permite identificar tanto puntos de convergencia entre ética y norma como vacíos estructurales donde la regulación jurídica resulta insuficiente frente a las exigencias éticas contemporáneas.

Convergencias: Alineación entre Ley Nacional y Estándares Éticos Internacionales

En términos generales, los marcos regulatorios nacionales sobre derechos del paciente, protección de datos personales y responsabilidad médica reflejan principios bioéticos ampliamente aceptados a nivel internacional, y esta convergencia demuestra que el derecho sanitario moderno no es ajeno a los fundamentos éticos de la práctica médica.

Protección de la Autonomía: La exigencia de consentimiento informado, presente en la mayoría de las legislaciones sanitarias nacionales, se encuentra alineada con el principio de autonomía defendido por la bioética clásica (Beauchamp & Childress, 2019). Asimismo, las directrices internacionales sobre IA en salud como las emitidas por la Organización Mundial de la Salud enfatizan la necesidad de transparencia y explicabilidad en el uso de sistemas algorítmicos, lo que fortalece la toma de decisiones informadas por parte del paciente (World Health Organization [WHO], 2021).

Principio de No Maleficencia y Seguridad del Paciente: Las normas nacionales que exigen certificación, evaluación técnica y vigilancia post-mercado de dispositivos médicos guardan coherencia con el principio de no maleficencia, y de igual forma, el enfoque de gestión del riesgo promovido por organismos como la Food and Drug Administration respecto al software como dispositivo médico (SaMD) se fundamenta en la prevención de daños previsibles (FDA, 2019).

Justicia y No Discriminación: La preocupación ética por la equidad en salud encuentra respaldo en normas antidiscriminatorias y en regulaciones de protección de datos que limitan el uso indebido de información sensible y a nivel internacional, el Parlamento Europeo ha establecido que los sistemas de IA de alto riesgo entre ellos los médicos deben cumplir requisitos estrictos de transparencia y mitigación de sesgos (European Commission, 2021).

Esta convergencia muestra que, al menos en su dimensión estructural, el derecho positivo incorpora progresivamente estándares éticos orientados a la justicia distributiva.

Vacíos Legales: Cuando la Ética Exige Más que la Norma

Pese a las convergencias señaladas, subsisten ámbitos donde la ética médica impone deberes que la ley difícilmente puede codificar de manera exhaustiva, estos vacíos no implican ausencia normativa absoluta, sino límites inherentes al derecho como sistema regulatorio.

La Empatía y la Dimensión Humana del Acto Médico

La ética médica tradicional reconoce la empatía como componente esencial del acto clínico, el encuentro médico-paciente no se reduce a una operación técnica de diagnóstico, sino que implica acompañamiento, comprensión emocional y construcción de confianza. Sin embargo, tales dimensiones relacionales no pueden ser reguladas con precisión jurídica sin caer en formulaciones excesivamente abstractas o simbólicas. La IA, al introducir automatización en procesos diagnósticos, puede optimizar precisión y eficiencia, pero no sustituye la dimensión interpersonal del cuidado. Topol (2019) sostiene que el reto no es reemplazar al médico, sino liberar tiempo clínico para fortalecer la relación humana. No obstante, la ley no puede obligar a "ser empático" de forma cuantificable, aunque la ética profesional sí lo exija como estándar moral.

Juicio Clínico Intuitivo y Prudencia Profesional

Otro vacío significativo se encuentra en el ámbito del juicio clínico intuitivo, entendido como la capacidad del profesional para integrar experiencia, contexto y particularidades del paciente más allá de los datos objetivos, donde los sistemas algorítmicos operan sobre patrones estadísticos, mientras que la ética médica reconoce el valor de la prudencia (phronesis) en la toma de decisiones complejas. La norma puede establecer deberes de diligencia, pero no puede

codificar exhaustivamente la prudencia clínica, cuando un médico decide apartarse de una recomendación algorítmica por considerar factores contextuales no capturados por el sistema, actúa dentro de un espacio ético que trasciende la regulación estricta y este margen interpretativo es indispensable para preservar la autonomía profesional

Un caso paradigmático surgió en Colombia, donde el Tribunal Superior de Bogotá anuló una sentencia condenatoria porque el juez de primera instancia empleó herramientas de inteligencia artificial para redactar la motivación del fallo, y en el proceso penal analizado, la sentencia original contenía citas jurisprudenciales y doctrinales inexistentes, generadas por la IA sin verificación humana, lo que vulneró el debido proceso del acusado (Tribunal Superior de Bogotá, 2 de diciembre de 2025). Los magistrados consideraron que la falta de supervisión humana implicó errores graves en la fundamentación jurídica y ordenaron la nulidad de la resolución y la elaboración de un nuevo fallo basado en fuentes verificables (Infobae, 2026; El Colombiano, 2025), lo que en este caso fue confirmado en una decisión posterior de la Corte Suprema de Justicia de Colombia, la cual advirtió que la información no verificada incorporada por la IA podía generar “alucinaciones” y causar defectos en la motivación judicial, afectando derechos fundamentales como el debido proceso (Infobae, 2026). Además, la Comisión de Disciplina Judicial inició investigaciones disciplinarias contra la jueza responsable, subrayando que el uso de IA sin supervisión puede generar responsabilidad profesional y sanciones incluso más allá del ámbito civil o penal (Infobae, 2025).

Estos precedentes son particularmente relevantes para la medicina porque ilustran cómo el sistema jurídico exige control humano efectivo sobre resultados automatizados y sanciona la delegación absoluta de decisiones en sistemas tecnológicos, una lógica que podría trasladarse al caso de fallos diagnósticos generados por IA médica.

Consentimiento Informado y Comunicación del Uso de IA en Diagnóstico Médico

El consentimiento informado es un requisito fundamental y obligatorio en la medicina moderna, basado en el respeto por la autonomía del paciente y consagrado como un derecho tanto ético como legal en muchas jurisdicciones, con estos términos generales, el consentimiento informado exige que el paciente reciba información adecuada, veraz y comprensible sobre los procedimientos, riesgos, beneficios y alternativas de una intervención sanitaria antes de aceptarla (Sentencia T-622, Corte Constitucional de Colombia, 2014; Sentencia T-1021, Corte Constitucional de Colombia, 2003).

¿Es obligatorio informar al paciente sobre el uso de IA?

Sí, aunque no existe una ley específica en muchos países que lo exija de forma explícita, tanto los marcos éticos internacionales como las interpretaciones jurídicas actuales apoyan la obligación de informar al paciente cuando se emplea IA en su diagnóstico o tratamiento, y esto se basa en los principios fundamentales de autonomía, transparencia y consentimiento informado. En casos donde el uso de IA representa un cambio significativo en la práctica clínica o un riesgo adicional, se recomienda informar de manera clara y, cuando corresponda, obtener un consentimiento explícito que incluya esa información como lo es el consentimiento informado clásico se aplica también a tecnologías nuevas cuando afectan decisiones de salud, las normas éticas internacionales como las de la OMS y la WMA sugieren comunicar el uso de IA en la atención médica, las regulaciones emergentes en Europa promueven transparencia de los sistemas de IA hacia las personas afectadas. La mejor práctica, aunque no universalmente legalmente codificada todavía, es informar al paciente de manera comprensible y significativa sobre el uso de IA en su diagnóstico o tratamiento clínico.

Estándares de "Explicabilidad" (XAI)

La incorporación de sistemas autónomos de inteligencia artificial en el diagnóstico médico ha reactivado un debate central: ¿qué nivel de explicabilidad debe ofrecer un algoritmo para que su resultado sea jurídicamente válido? La cuestión no es estrictamente técnica, sino profundamente normativa, en el ámbito sanitario, la validez de un diagnóstico no depende únicamente de su exactitud estadística, sino también de su trazabilidad, comprensibilidad y posibilidad de control.

La llamada Explainable Artificial Intelligence (XAI) surge como respuesta a la opacidad de los modelos complejos especialmente los basados en aprendizaje profundo cuya arquitectura interna no resulta intuitivamente interpretable y desde la perspectiva jurídica, la explicabilidad cumple tres funciones esenciales, lo que nos permite garantizar el derecho del paciente a la información, facilita la rendición de cuentas y atribución de responsabilidad, asegura la posibilidad de control judicial o pericial en caso de controversia. No obstante, las leyes no suelen exigir que el algoritmo sea totalmente transparente en términos de código fuente o fórmulas matemáticas, lo que demandan es un nivel de explicación suficiente para cumplir estándares de transparencia, seguridad y debido proceso**.**

Estándares en el Derecho Internacional y Europeo

El recientemente aprobado Unión Europea AI Act clasifica los sistemas de IA médica como “alto riesgo”, lo que implica exigencias estrictas de transparencia, documentación técnica, gestión de riesgos y supervisión humana (European Parliament & Council, 2024). Aunque el reglamento no obliga a revelar el algoritmo en su totalidad, sí exige la documentación detallada sobre el funcionamiento del sistema, registro de decisiones automatizadas, información clara para los usuarios profesionales, capacidad de supervisión humana efectiva, y, en consecuencia, la

explicabilidad exigida es funcional, no necesariamente estructural, donde es decir, el desarrollador debe demostrar cómo el sistema llega a resultados y bajo qué límites opera, pero no está obligado a divulgar secretos industriales si puede garantizar control y trazabilidad. Este estándar sugiere que, para ser legalmente válido, un diagnóstico apoyado por IA debe poder justificarse de manera comprensible ante autoridades regulatorias y tribunales, aunque no se exija que cada paciente entienda los detalles matemáticos del modelo.

Protección de Datos y Derecho a Explicación

El Parlamento Europeo, a través del Reglamento General de Protección de Datos (GDPR), reconoce el derecho del individuo a recibir "información significativa sobre la lógica aplicada" cuando una decisión se basa en procesamiento automatizado (GDPR, art. 22). Aunque existe debate doctrinal sobre si el GDPR establece un "derecho absoluto a explicación", sí se reconoce que las personas afectadas por decisiones automatizadas deben recibir una explicación comprensible de su impacto (Wachter et al., 2017). En el ámbito médico, esto implica que, si un diagnóstico depende sustancialmente de IA autónoma, el paciente podría tener derecho a conocer cómo influyó el sistema en la decisión.

Estándares en Estados Unidos

En Estados Unidos, la Food and Drug Administration (FDA) regula el software como dispositivo médico (SaMD), su enfoque se centra en el ciclo de vida del producto, validación clínica y control de modificaciones algorítmicas (FDA, 2019), donde la FDA no exige necesariamente explicabilidad total del modelo, pero sí requiere, tener presente la evidencia de desempeño clínico validado, documentación técnica suficiente para auditorías, gestión de riesgos y monitoreo post-mercado, ya que en este modelo, la validez legal del diagnóstico no depende de

que el algoritmo sea completamente interpretable, sino de que su seguridad y eficacia estén demostradas científicamente y exista capacidad de supervisión profesional.

En términos jurídicos, la exigencia es de explicabilidad razonable y proporcional al riesgo clínico y cuanto mayor sea el grado de autonomía del sistema y el impacto potencial en la vida del paciente, mayor será el nivel de transparencia requerido.

Los Sesgos Humanos y de IA

Sesgos Cognitivos Humanos y Sesgos Algorítmicos en Sistemas de Inteligencia Artificial Médica

La aparición de sistemas de inteligencia artificial en el diagnóstico clínico no elimina los errores asociados al proceso diagnóstico, sino que transforma su naturaleza al introducir una interacción compleja entre sesgos humanos y sesgos algorítmicos, y en este sentido, comprender cómo se originan y cómo interactúan ambos tipos de sesgos resulta fundamental para analizar los riesgos asociados al uso de sistemas autónomos de IA en medicina y su impacto en la seguridad del paciente.

Sesgos Cognitivos en el Razonamiento Clínico

El proceso de diagnóstico médico se fundamenta en el razonamiento clínico, entendido como un conjunto dinámico de procesos cognitivos mediante los cuales los profesionales de la salud interpretan la información clínica, generan hipótesis diagnósticas y toman decisiones terapéuticas. La evidencia más reciente en el campo de la seguridad del paciente y el razonamiento clínico ha reafirmado que este proceso no sigue una lógica completamente lineal ni puramente racional, sino que está influenciado por mecanismos cognitivos duales y por el uso de heurísticas. Estas estrategias mentales permiten tomar decisiones de forma eficiente en contextos de alta presión asistencial; sin embargo, diversos estudios posteriores a 2019 han demostrado que

también pueden favorecer la aparición de errores diagnósticos sistemáticos, especialmente en entornos complejos o con información incompleta (Graber et al., 2019; Norman et al., 2021).

Entre los sesgos cognitivos más frecuentes en el diagnóstico clínico, la literatura reciente continúa destacando el sesgo de anclaje, que se produce cuando el profesional se adhiere de manera excesiva a una hipótesis inicial y no la ajusta adecuadamente ante nueva evidencia clínica. De manera similar, el sesgo de disponibilidad ocurre cuando las decisiones diagnósticas se ven influenciadas por experiencias recientes o casos particularmente llamativos, lo que puede alterar la estimación real de la probabilidad de una enfermedad. Asimismo, el cierre prematuro sigue siendo reconocido como un factor crítico en el error diagnóstico, ya que implica la finalización anticipada del proceso de razonamiento sin considerar diagnósticos alternativos relevantes. Investigaciones contemporáneas han enfatizado que estos sesgos continúan siendo determinantes en la práctica clínica actual y representan un desafío importante para la mejora de la calidad y seguridad en la atención en salud (Singh et al., 2020; Saposnik et al., 2021).Estos sesgos no necesariamente reflejan falta de competencia profesional; más bien representan limitaciones inherentes al funcionamiento del sistema cognitivo humano, de hecho, el informe Improving Diagnosis in Health Care señala que los errores diagnósticos frecuentemente emergen de una combinación de factores cognitivos, organizacionales y sistémicos dentro de la práctica clínica (National Academies of Sciences, Engineering, and Medicine, 2015).

Sesgos Algorítmicos en Sistemas de Inteligencia Artificial

A diferencia de los sesgos humanos, los sesgos algorítmicos surgen de la forma en que los modelos de inteligencia artificial son diseñados, entrenados y aplicados, donde, los sistemas de aprendizaje automático dependen de grandes volúmenes de datos para identificar patrones y generar predicciones, pero, sin embargo, si los datos utilizados para entrenar el modelo contienen

desequilibrios o representaciones incompletas de ciertas poblaciones, el sistema puede reproducir y amplificar esas distorsiones en sus resultados (Obermeyer et al., 2019).

Un ejemplo significativo de sesgo algorítmico se observa cuando un modelo entrenado principalmente con datos de una población específica muestra un rendimiento inferior en otros grupos demográficos. Obermeyer et al. (2019) demostraron que un algoritmo utilizado en sistemas sanitarios estadounidenses subestimaba sistemáticamente las necesidades de salud de pacientes afroamericanos debido a la forma en que se construyó la variable objetivo utilizada en el entrenamiento del modelo.

Asimismo, los modelos de aprendizaje profundo pueden identificar correlaciones espurias presentes en los datos de entrenamiento. Zech et al. (2018) encontraron que algunos sistemas diseñados para detectar neumonía en radiografías torácicas aprendían a reconocer características institucionales de los hospitales en lugar de patrones clínicos relevantes, y este tipo de comportamiento revela cómo los algoritmos pueden desarrollar reglas de decisión incorrectas sin que los desarrolladores o los médicos lo detecten fácilmente.

La naturaleza compleja de muchos modelos de inteligencia artificial, especialmente aquellos basados en redes neuronales profundas, contribuye además a la dificultad de identificar estos sesgos. Como señala Rudin (2019), la opacidad de los modelos denominados de “caja negra” dificulta la auditoría de sus decisiones y complica la identificación de las causas específicas de un error diagnóstico.

Interacción entre Sesgos Humanos y Algorítmicos

En entornos clínicos donde médicos y sistemas de IA interactúan, los sesgos humanos y algorítmicos pueden reforzarse mutuamente, como, por ejemplo, cuando un profesional confía excesivamente en una recomendación automatizada, puede ignorar señales clínicas

contradictorias, fenómeno conocido como sesgo de automatización (Goddard et al., 2012). En estos casos, el error no se origina exclusivamente en el algoritmo ni en el médico, sino en la interacción entre ambos.

Por el contrario, también puede ocurrir el fenómeno opuesto, denominado aversión al algoritmo, donde los profesionales rechazan recomendaciones automatizadas incluso cuando estas presentan mayor precisión que la evaluación humana (Dietvorst et al., 2015). Ambas situaciones ilustran que la integración de la inteligencia artificial en la práctica clínica no elimina los sesgos existentes, sino que crea nuevas dinámicas de interacción que deben ser cuidadosamente gestionadas.

Relevancia para la Seguridad del Paciente

El reconocimiento de estos sesgos es fundamental para garantizar la seguridad del paciente en entornos clínicos asistidos por IA, y la literatura especializada coincide en que la implementación segura de estos sistemas requiere no solo mejorar el rendimiento técnico de los algoritmos, sino también diseñar mecanismos de supervisión humana, auditoría algorítmica y evaluación continua del desempeño en contextos clínicos reales (Kelly et al., 2019).

En consecuencia, la comprensión de los sesgos humanos y algorítmicos permite contextualizar el error diagnóstico en sistemas de inteligencia artificial como un fenómeno sociotécnico, en el que intervienen tanto factores cognitivos como tecnológicos, y este enfoque resulta esencial para desarrollar marcos regulatorios y estrategias de implementación que reduzcan los riesgos asociados al uso de sistemas autónomos de IA en medicina.

Marco Metodológico

La presente investigación se desarrolla bajo un enfoque cualitativo de tipo interpretativo, el cual permite analizar de manera integral fenómenos complejos relacionados con la implementación de sistemas de inteligencia artificial en el diagnóstico por imágenes médicas. A partir de la revisión documental, este enfoque facilita comprender no solo los aspectos técnicos, sino también las implicaciones éticas, clínicas y regulatorias asociadas a las vulnerabilidades en ciberseguridad y su impacto en los errores diagnósticos. Analizando la literatura científica reciente, se evidencia que los sistemas de inteligencia artificial en salud no pueden evaluarse únicamente desde una perspectiva cuantitativa, ya que su funcionamiento involucra múltiples dimensiones como la seguridad, la explicabilidad y la confiabilidad de los modelos, y, en este sentido, el enfoque cualitativo permite interpretar cómo factores como los ataques adversariales, los sesgos algorítmicos y la falta de transparencia influyen en la toma de decisiones clínicas (Topol, 2019; Geis et al., 2019).

El estudio adopta un diseño de investigación documental con alcance descriptivo y analítico, en el cual, a partir de la evidencia recopilada, se realiza una revisión sistematizada de fuentes secundarias con el fin de identificar patrones, vacíos de conocimiento y tendencias en el uso de la inteligencia artificial en el ámbito médico, especialmente en lo relacionado con la seguridad y confiabilidad de los sistemas (Sorin et al., 2023; Aggarwal et al., 2021). Dado que se trata de una investigación documental, la población está conformada por el conjunto de publicaciones científicas, documentos institucionales y marcos regulatorios relacionados con inteligencia artificial en salud, diagnóstico por imágenes y ciberseguridad.

La muestra se seleccionó de manera intencional, a partir de la revisión de literatura relevante y actualizada, en el cual, a partir de la evidencia encontrada, se priorizaron estudios

publicados entre los años 2019 y 2025, incluyendo artículos científicos, revisiones sistemáticas, documentos técnicos y normativas internacionales, y asimismo, se incluyeron publicaciones de organismos reconocidos como la Food and Drug Administration (FDA) y la Unión Europea, debido a su relevancia en la regulación de tecnologías médicas (European Commission, 2021; FDA, 2022).

Esta selección permitió garantizar que la información analizada fuera pertinente, confiable y alineada con los objetivos del estudio, centrados en comprender las vulnerabilidades de los sistemas de inteligencia artificial y su impacto en el diagnóstico médico, debido a la recolección de información que se realizó mediante la técnica de revisión documental, la cual constituye la principal herramienta en investigaciones de carácter cualitativo. A partir de la revisión bibliográfica, se identificaron y analizaron fuentes relevantes relacionadas con inteligencia artificial en diagnóstico médico, ataques adversariales, errores diagnósticos y marcos regulatorios.

Para ello, se emplearon bases de datos académicas reconocidas como PubMed, Scopus, ScienceDirect y Google Scholar, utilizando estrategias de búsqueda avanzada con palabras clave como “adversarial attacks in medical imaging”, “AI diagnostic errors”, “medical AI cybersecurity” y “AI regulation FDA EU”. Analizando la literatura disponible, se logró recopilar información suficiente para abordar la problemática desde diferentes perspectivas, y la información recolectada fue organizada y sistematizada mediante herramientas de gestión bibliográfica como Zotero, lo que permitió facilitar el registro, clasificación de la información. Además, se realizó una lectura crítica de cada documento, identificando conceptos clave y estableciendo relaciones entre los diferentes enfoques teóricos (Kitchenham et al., 2020).

Criterios de Inclusión y Exclusión

Para garantizar la calidad y pertinencia de la información, se establecieron criterios claros de inclusión y exclusión en el proceso de selección de fuentes.

Criterios de Inclusión

Publicaciones científicas entre 2019 y 2025

Estudios relacionados con inteligencia artificial en salud y diagnóstico por imágenes

Investigaciones sobre ciberseguridad, ataques adversariales y errores diagnósticos

Documentos regulatorios de organismos internacionales como la FDA y la Unión Europea

Artículos con respaldo académico y rigor metodológico

Criterios de Exclusión

Publicaciones sin respaldo científico o de fuentes no confiables

Estudios desactualizados o fuera del rango temporal establecido

Documentos que no abordan directamente la relación entre IA, diagnóstico médico y ciberseguridad

Información redundante o sin aporte significativo al análisis

A partir de estos criterios, se logró filtrar la información más relevante, permitiendo desarrollar un análisis sólido y coherente con los objetivos planteados.

El análisis de la información se llevó a cabo mediante la técnica de análisis de contenido, la cual permitió interpretar y organizar los datos obtenidos de la literatura revisada, y a partir de la evidencia recopilada, se identificaron categorías temáticas clave como: vulnerabilidades en sistemas de inteligencia artificial, ataques adversariales, errores diagnósticos (falsos positivos y falsos negativos), sesgos algorítmicos, explicabilidad de los modelos y marcos regulatorios.

Este proceso permitió establecer relaciones entre los diferentes conceptos y comprender cómo las fallas en ciberseguridad pueden traducirse en riesgos clínicos concretos, y a partir de la literatura reciente, se ha evidenciado que la integración entre tecnología, regulación y práctica clínica es fundamental para garantizar la seguridad del paciente y la confiabilidad de los sistemas de IA (Rajpurkar et al., 2022).

Desde el punto de vista ético, la investigación se fundamenta en el respeto por la propiedad intelectual, garantizando el uso adecuado. Donde, a partir de la revisión documental, se reconoce la importancia de abordar la inteligencia artificial en salud desde principios éticos como la transparencia, la responsabilidad y la protección de los datos del paciente. Analizando la evidencia disponible, diversos organismos internacionales destacan la necesidad de implementar sistemas de inteligencia artificial seguros, explicables y regulados, con el fin de minimizar riesgos y garantizar la calidad en la atención en salud (World Health Organization, 2021).

En coherencia con los fundamentos teóricos previamente expuestos, el presente trabajo se desarrolla bajo un enfoque metodológico de tipo descriptivo-analítico con alcance cualitativo, orientado a examinar de manera integral la relación entre la inteligencia artificial aplicada al diagnóstico por imágenes, los riesgos asociados a la ciberseguridad y la influencia de los sesgos cognitivos en el proceso diagnóstico. Para ello, se realizó una revisión narrativa de la literatura científica reciente, priorizando publicaciones a partir del año 2019 en bases de datos reconocidas como PubMed, Scopus y Web of Science, con el fin de garantizar la actualidad, pertinencia y validez de la información analizada.

El proceso metodológico incluyó la identificación, selección y análisis crítico de artículos científicos, guías internacionales y reportes técnicos relacionados con inteligencia artificial en salud, seguridad de la información, errores diagnósticos y sesgos cognitivos. Como criterios de

inclusión se consideraron estudios en idioma inglés y español, con enfoque en aplicaciones clínicas de la IA, vulnerabilidades en sistemas médicos digitales y factores humanos que influyen en la toma de decisiones. Se excluyeron aquellos documentos sin respaldo académico o con información desactualizada, asegurando así la calidad de las fuentes utilizadas.

Posteriormente, la información recopilada fue organizada en categorías temáticas que permitieron estructurar el marco teórico: fundamentos de la inteligencia artificial en imágenes médicas, ciberseguridad en sistemas de salud, vulnerabilidades tecnológicas y sesgos cognitivos en el razonamiento clínico. Este proceso facilitó la integración de los distintos conceptos, permitiendo establecer relaciones entre los factores tecnológicos y humanos que intervienen en la seguridad del paciente.

Finalmente, el análisis se orientó a una interpretación crítica de la evidencia, con el propósito de identificar riesgos, limitaciones y oportunidades de mejora en el uso de estas tecnologías en el entorno clínico. Este enfoque metodológico permite no solo comprender el estado actual del conocimiento, sino también sustentar la necesidad de implementar estrategias que fortalezcan la seguridad, confiabilidad y uso ético de la inteligencia artificial en el ámbito de la salud.

Desarrollo del Proyecto

El avance de la inteligencia artificial (IA) en el ámbito de la salud ha generado una transformación significativa en el diagnóstico por imágenes, mejorando la precisión, la rapidez y la eficiencia en la detección de diversas patologías. En particular, los modelos de aprendizaje profundo han demostrado un desempeño comparable e incluso superior al de especialistas humanos en tareas específicas como la identificación de lesiones, tumores y anomalías. Además, su implementación ha permitido optimizar los flujos de trabajo clínicos mediante la automatización de procesos, la priorización de casos urgentes y la reducción de los tiempos de respuesta en contextos de alta demanda asistencial (Hosny et al., 2018; Topol, 2019; Esteva et al., 2017).

No obstante, el análisis evidencia que el rendimiento de estos sistemas depende en gran medida de la calidad, diversidad y representatividad de los datos utilizados durante su entrenamiento, y en este sentido, una de las principales problemáticas identificadas corresponde a las vulnerabilidades técnicas, entre las que destacan los ataques adversariales y el envenenamiento de datos. Estas amenazas pueden generar alteraciones imperceptibles en las imágenes médicas o introducir sesgos en los datos, induciendo errores diagnósticos sin que sean fácilmente detectables, lo que compromete directamente la seguridad del paciente (Finlayson et al., 2019; Paschali et al., 2021).

En relación con los errores diagnósticos, se identifican principalmente los falsos positivos y los falsos negativos, ya que los primeros pueden derivar en intervenciones innecesarias, aumento de costos y afectación emocional del paciente, mientras que los segundos representan un riesgo mayor al implicar la omisión de una patología real, retrasando el tratamiento y empeorando el pronóstico clínico, y estos errores no solo dependen del desempeño técnico del

modelo, sino también de factores como los sesgos algorítmicos y la interacción con el profesional de la salud (Obermeyer et al., 2019; Kelly et al., 2019).

Otro aspecto crítico es la falta de transparencia en los sistemas de IA, conocida como el fenómeno de la “caja negra”, que dificulta comprender el proceso mediante el cual se generan las decisiones diagnósticas. Esta opacidad limita la validación clínica, reduce la confianza del personal médico y complica la identificación de fallos o la atribución de responsabilidades, y asimismo, los sesgos en los datos de entrenamiento pueden generar desigualdades en la calidad del diagnóstico, afectando de manera desproporcionada a ciertos grupos poblacionales (European Commission, 2021; WHO, 2021).

Desde el punto de vista regulatorio, organismos internacionales han comenzado a establecer lineamientos para garantizar la seguridad y eficacia de estos sistemas. Tanto la Food and Drug Administration (FDA) como la Unión Europea coinciden en la necesidad de aplicar enfoques basados en el riesgo, la evaluación continua y la supervisión humana, considerando la IA como una herramienta de apoyo y no como un sustituto del profesional de la salud. En este contexto, se promueve la implementación de mecanismos de auditoría, monitoreo y gestión de riesgos (FDA, 2019; European Commission, 2021).

Finalmente, se destaca que la interacción entre el ser humano y la inteligencia artificial constituye un elemento determinante en la seguridad del paciente. Fenómenos como el sesgo de automatización y la aversión al algoritmo evidencian que el error diagnóstico puede surgir de la relación entre ambos, y por ello, se hace necesario fortalecer la formación del personal sanitario, mejorar la explicabilidad de los sistemas y garantizar un equilibrio adecuado entre confianza y supervisión (Cabitza et al., 2017; Topol, 2019).

En síntesis, aunque la inteligencia artificial representa una herramienta poderosa para el diagnóstico médico, su implementación segura requiere un enfoque integral que combine avances tecnológicos, ciberseguridad, regulación robusta y supervisión humana constante, con el fin de garantizar un uso ético, confiable y responsable en los entornos de salud (WHO, 2021; European Commission, 2021).

Tabla 3.

Comparación de hallazgos sobre IA y vulnerabilidades en la seguridad diagnóstica.

Autor(es) (Año)	Enfoque principal	Hallazgo clave	Implicación en seguridad diagnóstica
Topol (2019)	IA en medicina clínica	La IA mejora la precisión diagnóstica mediante análisis de grandes volúmenes de datos	Requiere supervisión humana para evitar dependencia excesiva y errores no detectados
Finlayson et al. (2019)	Ciberseguridad en IA médica	Los modelos son vulnerables a ataques adversariales que alteran imágenes sin ser perceptibles	Riesgo directo de diagnósticos erróneos y amenaza a la seguridad del paciente
Rudin (2019)	Explicabilidad de modelos	Los sistemas tipo “caja negra” limitan la comprensión del proceso de decisión	Dificulta la validación clínica y la atribución de responsabilidades
Obermeyer et al. (2019)	Sesgos algorítmicos	Los algoritmos pueden reproducir desigualdades por datos no representativos	Genera errores sistemáticos y afecta la equidad en la atención
Kelly et al. (2019)	Implementación clínica de IA	La efectividad depende de validación en entornos reales y calidad de datos	Riesgo de bajo rendimiento si no se adapta al contexto clínico
Cabitza et al. (2017)	Interacción humano-IA	Existe sesgo de automatización y dependencia del sistema	Puede reducir el juicio crítico del profesional
Rajkomar et		La calidad y cantidad de datos	Datos deficientes aumentan la

al. (2019)	IA y datos clínicos	determinan el desempeño del modelo	probabilidad de error diagnóstico

Autor(es) (Año)	Enfoque principal	Hallazgo clave	Implicación en seguridad diagnóstica
WHO (2021)	Ética y gobernanza	La IA debe garantizar transparencia, equidad y seguridad	La protección del paciente es un principio central en la implementación

Nota. Elaboración propia con base en la revisión de literatura sobre Comparación de hallazgos sobre IA y vulnerabilidades en la seguridad diagnóstica.

Conclusiones

A partir de la revisión documental y del análisis de la evidencia científica, se concluye que la relación entre ciberseguridad y seguridad del paciente trasciende el ámbito técnico y se configura como un componente esencial de la ética en salud. La evidencia revisada permite afirmar que la protección de los sistemas de inteligencia artificial no solo busca resguardar datos clínicos, sino garantizar la integridad del proceso diagnóstico. En este sentido, la posibilidad de manipulación de imágenes médicas o alteración de algoritmos mediante ataques adversariales representa un riesgo directo para la vida del paciente, ya que puede conducir a decisiones clínicas erróneas sin que el profesional lo advierta. Por lo tanto, la ciberseguridad debe entenderse como un pilar fundamental en la prevención del error diagnóstico en entornos digitales, alineado con el principio de no maleficencia y la obligación de garantizar una atención segura (Finlayson et al., 2019; WHO, 2021).

Asimismo, los hallazgos derivados del análisis de la literatura permiten sostener que la implementación de la inteligencia artificial en el diagnóstico médico debe desarrollarse bajo un modelo híbrido, en el cual la tecnología actúe como herramienta de apoyo y no como sustituto absoluto del juicio clínico. El enfoque human-in-the-loop se consolida como una condición necesaria para preservar la responsabilidad profesional, ya que asegura la participación del médico en la validación de los resultados generados por el sistema. La evidencia muestra que la ausencia de supervisión humana significativa puede incrementar el riesgo de errores no detectados, especialmente en sistemas autónomos. En consecuencia, la interacción entre el profesional y la IA debe concebirse como un proceso colaborativo que combine la capacidad analítica de los algoritmos con la experiencia clínica y el juicio prudencial del médico (Topol, 2019; European Commission, 2021).

Por otra parte, a partir de la evidencia revisada, se destaca la importancia de la explicabilidad como requisito indispensable para la implementación segura de la inteligencia artificial en salud. La naturaleza de "caja negra" de muchos modelos limita la comprensión de sus decisiones, lo que dificulta la validación clínica, la confianza del profesional y la atribución de responsabilidades en caso de error. En este contexto, la literatura enfatiza que no es suficiente alcanzar altos niveles de precisión estadística; es necesario que los sistemas permitan una interpretación razonable de sus resultados, de modo que el profesional pueda integrar la información algorítmica dentro de un proceso de razonamiento clínico informado. La explicabilidad, por tanto, se vincula directamente con principios éticos como la autonomía, la transparencia y la rendición de cuentas (Rudin, 2019; Samek et al., 2021).

En relación con el marco normativo, el análisis documental evidencia la necesidad urgente de fortalecer y actualizar las regulaciones existentes para responder a los desafíos que plantea la inteligencia artificial en el ámbito sanitario. La experiencia internacional, particularmente el enfoque basado en riesgo adoptado por la Unión Europea y las directrices de la Organización Mundial de la Salud, sugiere que los sistemas de IA médica deben ser clasificados como tecnologías de alto riesgo, lo que implica la exigencia de estándares rigurosos de validación, supervisión y vigilancia postcomercialización. En el contexto colombiano, aunque existen normas aplicables en materia de salud, dispositivos médicos y protección de datos, se identifica la necesidad de desarrollar marcos específicos que integren la dimensión tecnológica con la protección de los derechos del paciente. Esto permitiría garantizar un uso seguro, equitativo y transparente de estas herramientas en la práctica clínica (European Commission, 2021; WHO, 2021).

Finalmente, la evidencia analizada permite concluir que la incorporación de la inteligencia artificial en la medicina implica una transformación en el perfil profesional de los trabajadores de la salud, particularmente en áreas como la radiología. Ya no es suficiente contar únicamente con conocimientos clínicos tradicionales, sino que se requiere una formación complementaria que permita comprender el funcionamiento, las limitaciones y los posibles sesgos de los sistemas algorítmicos. Esta competencia técnica se integra al deber de cuidado del profesional, ya que influye directamente en su capacidad para interpretar, validar y cuestionar los resultados generados por la IA. En este sentido, la educación médica debe adaptarse a los nuevos entornos digitales, promoviendo un enfoque interdisciplinario que combine conocimientos clínicos, tecnológicos y éticos (Benjamens et al., 2020; Kelly et al., 2019).

En conjunto, estas conclusiones, sustentadas en la revisión de la evidencia científica, reflejan que la inteligencia artificial en el diagnóstico médico no solo representa un avance tecnológico, sino un cambio estructural en la forma de entender la seguridad del paciente, la responsabilidad profesional y la práctica clínica contemporánea. Su implementación segura dependerá de la capacidad de integrar adecuadamente la innovación tecnológica con principios éticos, marcos regulatorios sólidos y una formación profesional acorde a los nuevos desafíos del entorno médico.

Referencias

American College of Radiology. (2024). *Data workflow efficiency standards*. American College of Radiology.

American College of Radiology. (2024). *Ethics of artificial intelligence in radiology: Summary of the joint statement*. American College of Radiology.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning. *Nature Medicine*, *25*, 1301–1309. https://doi.org/10.1038/s41591-019-0508-1

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care. *The New England Journal of Medicine*, *378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229

European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. European Commission.

Food and Drug Administration. (2019). *Proposed regulatory framework for modifications to artificial intelligence/machine learning-based software as a medical device*. FDA document.

Food and Drug Administration. (2021). *Artificial intelligence/machine learning (AI/ML)-based software as a medical device action plan*. FDA document.

Food and Drug Administration. (2024). *Regulatory framework for AI/ML-based medical device software*. U.S. Food and Drug Administration.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*, 30–36. https://doi.org/10.1038/s41591-018-0307-0

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial

intelligence in radiology. *Nature Reviews Cancer*, *18*, 500–510. https://doi.org/10.1038/s41568-018-0016-5

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*, 230–243. https://doi.org/10.1136/svn-2017-000101

Journal of Digital Imaging. (2024). The role of AI in transforming teleradiology workflows. *Springer Journal of Digital Imaging*.

Journal of the American Medical Association. (2024). Comparative studies on AI vs. human performance. *JAMA*.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, *17*, 195. https://doi.org/10.1186/s12916-019-1426-2

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care. *The Lancet Digital Health*, *2*(5), e272–e279. https://doi.org/10.1016/S2589-7500(20)30014-1

National Institute of Standards and Technology. (2024). *Adversarial machine learning in healthcare*. NIST.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm

used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Organización Mundial de la Salud. (2021). *Ethics and governance of artificial intelligence for health*. WHO publication.

Organización Mundial de la Salud. (2024). *Generating evidence for artificial intelligence-based medical devices*. WHO.

Price, W. N., Gerke, S., & Cohen, I. G. (2019). Potential liability for physicians using artificial intelligence. *JAMA*, *322*(18), 1765–1766. https://doi.org/10.1001/jama.2019.15064

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (2021). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer. https://doi.org/10.1007/978-3-030-28954-6

Sociedad Española de Radiología Médica. (2024). *Informe sobre la integración de la IA en el flujo de trabajo radiológico*. SERAM.

Sorin, V., Barash, Y., Konen, E., & Klang, E. (2020). Deep learning for medical imaging-based cancer diagnosis. *The Lancet Oncology*, *21*(8), e367–e378. https://doi.org/10.1016/S1470-2045(20)30108-1

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems. *NPJ Digital Medicine*, *3*, 17. https://doi.org/10.1038/s41746-020-0221-y

Topol, E. (2019). High-performance medicine: The convergence of human and artificial intelligence.

Nature Medicine, *25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7

Topol, E. J. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. *PLoS Medicine*, *15*(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683